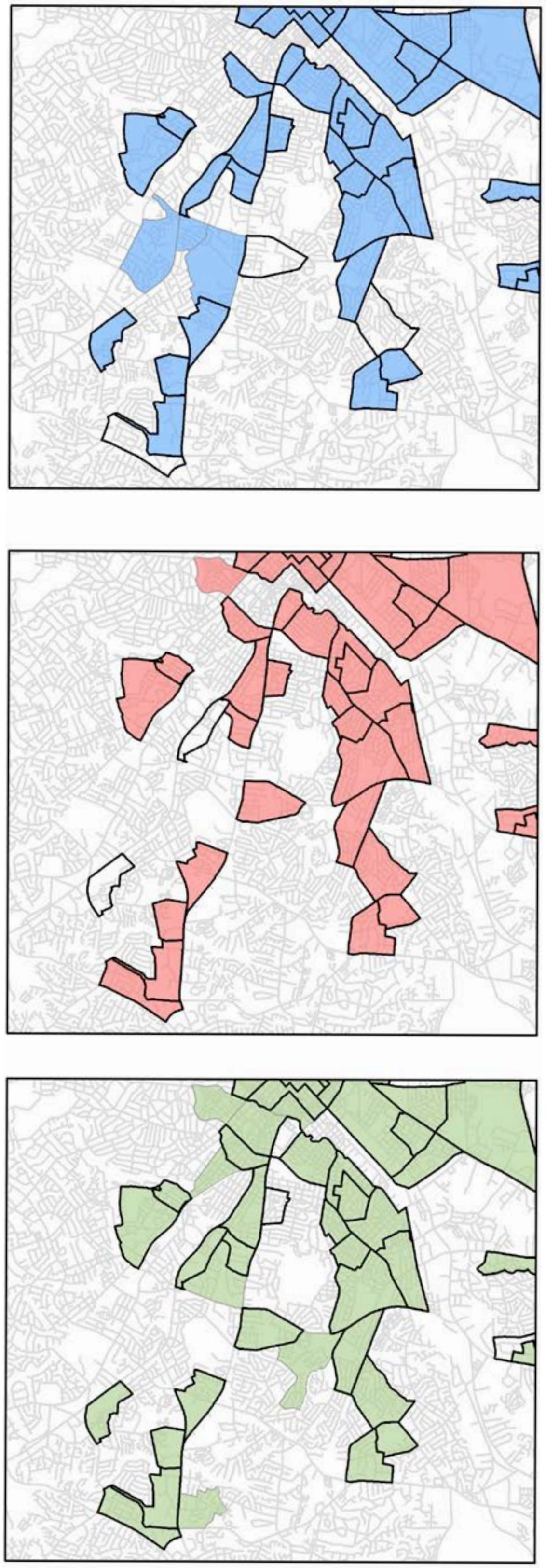
Document Title:	Effects of Data Quality on Predictive Hotspot Mapping
Author:	Timothy C. Hart, Ph.D., Paul A. Zandbergen, Ph.D.
Document No.:	239861
Date Received:	October 2012
Award Number:	2009-IJ-CX-0022



Effects of Data Quality on Predictive Hotspot Mapping

FINAL TECHNICAL REPORT

National Institute of Justice
Cooperative Agreement # 2009-IJ-CX-0022

September 1, 2012

Prepared for

Joel Hunt
Mapping and Analysis for Public Safety (MAPS) Program
Office of Research and Evaluation
National Institute of Justice
810 7th St., NW
Washington, DC 20531

Prepared by

Timothy C. Hart, Ph.D.
Department of Criminal Justice
University of Nevada, Las Vegas
4505 S. Maryland Pkwy.
Box 5009
Las Vegas, NV 89154

AND

Paul A. Zandbergen, Ph.D.
Department of Geography
Bandelier West Room 111
MSC01 1110
1 University of New Mexico
Albuquerque, NM 87131

EXECUTIVE SUMMARY

The purpose of the current research was to contribute to the improved robustness of predictive crime mapping techniques. Our goal was to investigate the effect of data quality on predictive hotspot mapping analysis in order to achieve the following three objectives:

1. Determine empirical descriptions of the quality of a range of “typical” geocoding techniques employed in crime mapping, including their completeness, positional accuracy and repeatability;

2. Characterize the effects of data quality on the robustness of selected predictive crime hotspot mapping techniques; and

3. Determine the effects of analysis method, crime type, urban morphology and parameter settings for predictive crime hotspot mapping techniques given a range of typical data quality parameters within the context of the accuracy and precision of hotspot prediction.

The current study analyzed over 400,000 crime incident records from six large law enforcement jurisdictions in the U.S.

Geocoding Quality Analysis

Collectively, results from the current study suggest that geocoding quality is affected by variations in crime type as well as reference data used during the geocoding process. In order to increase the completeness and positional accuracy of street geocoded crime events, we developed five general recommendations. Based on our findings, when geocoding crime data is a necessary part of their research and address point or parcel reference data are unavailable, analysts and researchers should…

1. Assess the overall quality of input address information prior to geocoding

2. Disaggregate crime incidents and geocode like crime events separately

3. Tailor geocoding procedures to fit specific needs

4. Geocode to local street centerline reference data, if it is available

5. Characterize positional accuracy prior to additional analysis

We also conclude that future research in this area should focus on ways to improve the overall quality of input address information for crime events, especially for those types of crimes whose geocoding quality is more adversely impacted during the geocoding process by the overall quality of the input address information (i.e., burglary). In addition, we feel future research should examine how the manipulation of user-defined parameter settings contained within the

address locator service influences geocoding quality. Finally, we suggest that future research should consider how variations in geocoding quality impact various crime analysis techniques such as trend analysis.

Predictive Hotspot Analysis

Based on our predictive hotspot analysis, we determined that no single technique is more accurate than any other. Instead, these procedures are highly influenced by many factors. Nevertheless, a series of recommendations based on our results are presented below and offer guidance for analysts and researchers engaging in hotspot analysis. When conducting hotspot analysis, we recommend that analysts and researchers…

1. Consider analyzing data with multiple techniques
2. Disaggregate crime incidents and analyze like crime events separately
3. Take study area into consideration
4. Be cognizant of user-defined parameter settings
5. Use street centerline reference data or address point reference data
6. Determine how predictive accuracy will be measured

As with any research, the current study is not without certain limitation. For example, data used was limited to six geographic areas. Although attempts were made to include agencies that represent a mix of urban and rural locations, specific match rates and positional error results may be unique to these study areas. Additionally, effects of changes in user-defined parameters considered during the geocoding process (e.g., spelling sensitivity, minimum candidate score, and the exclusion of tied candidates) and the influence of variations in matching algorithms of different geocoding software were not investigated. And since every crime incident could not be linked to a corresponding address point/parcel location, estimates of positional error are likely biased downward since the general quality of the records not geocoded is likely lower.

In terms of the direction of future predictive crime hotspot research, we determined that it should focus on how measures of predictive accuracy can be enhanced as well as explore ways to enhance traditional output of hotspot analysis.

Timothy C. Hart, Ph.D.
Director, Center for the Analysis of Crime Statistics

TABLE OF CONTENTS

EXECUTIVE SUMMARY .... **ii**
TABLE OF CONTENTS .... **iv**
LIST OF FIGURES .... **vi**
LIST OF TABLES .... **vii**
1. ABSTRACT **1**
2. INTRODUCTION **2**
3. LITERATURE REVIEW **3**
3.1. Introduction on Crime Mapping and Analysis 3
3.2. Types of Crime Analysis 4
3.2.1. *Administrative crime analysis* 4
3.2.2. *Tactical crime analysis* 4
3.2.3. *Strategic crime analysis* 4
3.3. Crime Hotspot Analysis 5
3.4. Predictive Crime Mapping 6
3.5. Background on Geocoding 7
3.6. Geocoding Quality 11
3.7. Effects of Geocoding Quality on Spatial Analysis 13
4. GEOCODING QUALITY ANALYSIS **15**
4.1. Research Questions 15
4.2. Data and Methodology 15
4.3. Results 19
4.3.1. *Match rates* 19
4.3.2. *Positional accuracy* 21
4.4. Discussion and Conclusions 23
4.5. Limitations 26
4.6. Future Research 26
4.7. Recommendations 27
5. PREDICTIVE HOTSPOT MAPPING ANALYSIS **29**
5.1. Research Questions 29
5.2. Data and Methodology 29
5.2.1. *Measures of predictive accuracy* 30
5.2.2. *Hotspot techniques* 30
5.3. Results 32
5.3.1. *Hotspot method* 32
5.3.2. *Crime type* 32
5.3.3. *Study area* 36
5.3.4. *Parameter settings* 38
5.3.5. *Geocoding quality* 46
5.4. Discussion and Conclusions 52
5.5. Limitations 54
5.6. Future Research 54

5.7. Recommendations 55

6. REFERENCES 57
7. DISSEMINATION OF INFORMATION 61
7.1. Conference Presentations 61
7.2. Journal Article Submissions/Publications 61
7.3. Workshops 62

APPENDIX A 63
APPENDIX B 69
APPENDIX C 71

LIST OF FIGURES

Figure 1. Illustration of the geocoding process, whereby a latitude and longitude (i.e., XY coordinate) is assigned to an input address (left), based on the known location of reference address information (right)....................................7
Figure 2. Elements of an address used in the geocoding process. 8
Figure 3. Conceptual diagram of the algorithm behind street geocoding. 10
Figure 4. Example of address points and parcel boundaries for single-family residential area in Henderson, Nevada. 10
Figure 5. Illustration of an input address geocoded to three different reference layers 11
Figure 6. Positional accuracy illustrated. 12
Figure 7. Example of a Geocoding Options interface where a user can manipulate settings such as spelling sensitivity, minimum candidate scores, minimum match scores, offsets, etc. 18
Figure 8. Aggregated and point pattern analysis methods and corresponding parameter settings that were used in the current study to examine the effects of geocoding quality on predictive hotspot mapping 30
Figure 9. Examples of how the hit rate, Predictive Accuracy Index (PAI) and Recapture Rate Index (RRI) are calculated. 31
Figure 10. An illustration of how variation in the parameter setting that defines a “cluster” affects predictive accuracy (i.e., hit rate (HR), the Predictive Accuracy Index (PAI), and the Recapture Rate Index (RRI))............................31
Figure 11. Predictive accuracy metrics produced from KDE. Analysis based on assaults in Charlotte-Mecklenburg, using different threshold sizes and search radii 39
Figure 12. Predictive accuracy metrics produced from KDE. Analysis based on auto burglaries in Charlotte-Mecklenburg, using different threshold sizes and search radii 40
Figure 13. Predictive accuracy metrics produced from KDE. Analysis based on auto thefts in Charlotte-Mecklenburg, using different threshold sizes and search radii 41
Figure 14. Predictive accuracy metrics produced from KDE. Analysis based on burglaries in Charlotte-Mecklenburg, using different threshold sizes and search radii 42
Figure 15. Predictive accuracy metrics produced from KDE. Analysis based on drug offenses in Charlotte-Mecklenburg, using different threshold sizes and search radii 43
Figure 16. Predictive accuracy metrics produced from KDE. Analysis based on homicides in Charlotte-Mecklenburg, using different threshold sizes and search radii 44
Figure 17. Predictive accuracy metrics produced from KDE. Analysis based on robberies in Charlotte-Mecklenburg, using different threshold sizes and search radii 45
Figure 18. Variations in two key parameter settings used to produce KDE maps............53

LIST OF TABLES

Table 1. Crime incidents by crime type and jurisdiction, 2007-08 17
Table 2. Geocoding match rate results by crime type and type of street reference data, 2007-08 20
Table 3. Positional error statistics for street geocoded crime events in all jurisdictions, 2007-08 22
Table 4. The influence of hotspot method on predictive accuracy for street geocoded crimes within all jurisdictions 33
Table 5. The influence of crime type on predictive accuracy for street geocoded incidents, by hotspot method 34-35
Table 6. The influence of urban morphology on predictive accuracy for street geocoded incidents, by hotspot method 36-37
Table 7. The influence of cell size and geocoding method on the predictive accuracy of grid-based thematic maps of assaults and burglaries, by jurisdiction 46-47
Table 8. The influence of cell size and geocoding method on the predictive accuracy of Local Moran's I maps of assaults and burglaries, by jurisdiction 48
Table 9. The influence of cell size and geocoding method on the predictive accuracy of Getis-Ord Gi* maps of assaults and burglaries, by jurisdiction 49-50
Table 10. The influence of cell size and geocoding method on the predictive accuracy of KDE maps of assaults and burglaries, by jurisdiction 51

1. ABSTRACT

A growing interest and use of crime mapping and analysis among practitioners and academics alike, the potential impact that geocoding results can have on the spatio-temporal analysis of crime, and a sparse literature in this area gave rise to the current study, which was divided into two complementary components.

The aim of the first component of our research was to investigate the relationship between reference data and geocoding quality for crime analysis, whereas the second component of our research was designed to determine the impact of data quality on predictive hotspot mapping techniques. Approximately 400,000 crime incident records from six large law enforcement jurisdictions in the U.S. were analyzed: Arlington Police Department (TX), Albuquerque Police Department (NM), Charlotte-Mecklenburg Police Department (NC), Las Vegas Metropolitan Police Department (NV), San Diego County Sheriff's Office (CA), and Tampa Police Department (FL).

Completeness and positional accuracy of geocoding results was assessed across several crime types as well as different levels of input address quality, using both commercial and non-commercial street network reference data. Results suggest that 1) the type of reference data (e.g., streets, parcel centroids, or address points) used in the geocoding process affects geocoding quality; 2) match rates vary by crime type and are influenced greatly by the quality of input addresses; 3) the type of reference data in conjunction with crime type influences geocoding quality considerably; 4) the type of reference data used in the geocoding process affects geocoding quality, measured in terms of positional accuracy; and 5) positional accuracy of geocoded crime incidents is influenced greatly by crime type.

Predictive accuracy of hotspot maps was based on three metrics, including hit rate, Predictive Accuracy Index (PAI), and Recapture Rate Index (RRI). Findings show that the effect of geocoding quality on predictive hotspot crime mapping varies by crime type, urban morphology, and technique, as well as parameter settings associated with them. More importantly, findings suggest that the effect of geocoding quality on predictive hotspot mapping is complex and often mitigated by many of the aforementioned factors. In sum, findings suggest that no one predictive hotspot mapping technique is superior to any other.

The study's contribution to the literature, limitations, and implications for future research are discussed. More importantly, specific recommendations related to both geocoding and hotspot analyses are offered.

2. INTRODUCTION

Predicting the location and time of future crime events is of great interest to law enforcement. One approach in crime prediction is the use of crime hotspots, which rely on the assumption that the locations of past events are good predictors of future events. Several approaches have been developed recently to assess the performance of hotspot analysis techniques for crime prediction, including the Prediction Accuracy Index (PAI) and the Recapture Rate Index (RRI). Most types of crime analysis, including predictive hotspot mapping, rely heavily on geocoded crime locations, but errors in geocoding can be substantial.

The widespread availability of powerful geocoding tools in commercial GIS software and the interest in spatial analysis at the individual level has made address geocoding a widely employed technique in many different fields, including criminology. Many crime analysis techniques rely on the ability to geocode crime incidents based on address information. The quality of geocoding and its effect on spatial analysis have received some attention in the literature, in particular in the health field. Crime incident data, however, is somewhat unique in its characteristics and so are crime analysis techniques that employ geocoded crime incidents.

The effect of geocoding quality on the robustness of spatial analysis of crime incidents has received limited attention. These effects need to be properly characterized to increase our confidence in the results of predictive hotspot mapping. This analysis needs to be conducted within the broader context of determining the robustness of hotspot mapping for crime prediction.

The current study was designed to contribute to our overall understanding of predictive crime mapping techniques by investigating the effects of geocoding quality of predictive hotspot mapping. Specifically, the current study determined empirical descriptions of the quality of a range of "typical" geocoding techniques employed in crime mapping, including their completeness, positional accuracy and repeatability[1]. Furthermore, the study characterized the effects of data quality on the robustness of selected predictive crime hotspot mapping techniques; and examines the influence of various parameters settings for different predictive crime hotspot mapping procedures, given a range of typical data quality parameters within the context of the accuracy and precision of hotspot prediction.

The remainder of this report is organized in the following manner: First, an overview of the relevant literature is offered, focusing on the geocoding process, common ways to geocode data, how the quality of geocoded information is traditionally determined, and the effects of geocoding on spatial analysis. The next two sections contain our research questions, descriptions of the data and methods, analytic strategies, and results of our analyses for both components of this study (i.e., the geocoding analysis and the predictive hotspot analysis), respectively. Each of these sections conclude with a discussion of the implications of our findings, limitations of the study,

[1] As will be described below in greater detail, effects of user-defined parameter settings used during the geocoding process were not examined as part of the current study. Rather, settings were held constant across all procedures.

suggestions for future research, and specific recommendations to analysts and academics alike. We begin the report with a review of the relevant literature.

3. LITERATURE REVIEW

3.1. Introduction on Crime Mapping and Analysis

Law enforcement agencies throughout the country have employed Geographic Information Systems (GIS) for combating crime. Results from a recent Bureau of Justice Statistics (BJS) survey show that more than 87% of the nation's largest law enforcement agencies have personnel designated specifically to crime analysis duties, and that more than a third of large agencies provide direct access to crime maps produced by crime analysts via the web (personal communication with Brian Reaves, BJS Statistician, on January 21, 2011). Other uses of GIS among law enforcement analysts include monitoring and tracking sex offenders, geographic profiling, and identifying crime hot spots (Reaves & Hart, 2000). In sort, law enforcement agencies across the United States have identified GIS as a valuable tool for addressing the myriad facets of fighting crime (Mamalian & LaVigne, 1999; Weisburd & Lum, 2005).

In conjunction with the increased use of crime mapping and analysis among law enforcement agencies, a growing number of applied and scientific publications examining the use and role of GIS with the field of criminology have been published. For example, some of the leading journals in the field of criminology such as the *Journal of Quantitative Criminology, Criminology, Criminology,* and the *British Journal of Criminology* have release a number of articles on crime mapping and analysis over the past several years (e.g., Andresen, 2006; Anselin, Cohen, Cook, Gorr, & Tita, 2000; Bernasco & Nieuwbeerta, 2005; Bowers & Johnson, 2003; Bowers et al., 2004; Grubesic, 2006; Murray et al., 2001; Poulsen & Kennedy, 2004; Ratcliffe & McCullagh, 2001; Ratcliffe, 2005; Wang, 2005a; Zandbergen & Hart, 2006, 2009). Similarly, the publication of prominent reports and recently published books also suggest strongly that crime mapping and analysis have come of age (e.g., Boba, 2013; Chainey & Ratcliffe, 2005; Chainey & Thompson, 2008a; Eck, Chainey, Cameron, Leitner, & Wilson, 2005; Hirschfield, 2007; LaVigne, 2007; Lersch & Hart, 2011; Paulsen & Robinson, 2008; Wang, 2005b).

Combined, this growing body of literature suggests that crime mapping and analysis has strong theoretical grounding (including social ecology theories and place-base theories), well-developed analytic methodologies (descriptive analysis, hot-spot detection, explanatory regression analysis, spatial modeling, etc.), and widespread implementation (geographic profiling, resource allocation, crime prevention, etc.).

Unfortunately, the speed at which GIS is being embraced and utilized by both academics and practitioners is outpacing the scientific literature on some of the important methodological issues associated with this powerful analytic approach. For example, to date little is known about the quality of crime incident data produced from techniques that create a feature on a map from the address location of a crime incident—a process known as geocoding—and the consequences of data quality on predictive hotspot mapping. In response, the current study was designed to fill this gap in the literature by 1) examining the influence of reference data quality on the completeness and positional accuracy of street geocoded crime data; and 2) demonstrating the effects of geocoding quality on an array of predictive hotspot mapping techniques. Information produced from the current study provides researchers and practitioners alike with general guidelines to follow in order to 1) maximize the quality of the data they commonly use in their spatio-temporal analyses of crime; and to 2) effectively and efficiently use this information to predict future crime events.

3.2. Types of Crime Analysis

Crime analysis and mapping can take many forms. Boba (2012) suggest that each can be classified into one of three types of analytic techniques: Administrative Crime Analysis (ACA), Tactical Crime Analysis (TCA), and Strategic Crime Analysis (SCA). Each of these approaches is summarized in the following section.

3.2.1. *Administrative crime analysis*. Administrative Crime Analysis (ACA) typically involves long range projects, often internal to the agency. Common practices associated with ACA include providing economic, geographic and law enforcement information to police management, city hall, city council, and neighborhood, citizen groups, or the media. While results produced from ACA are often the same results that are produced from the other analytic approaches, Boba suggests that information chosen for presentation in ACA "represents only the 'tip of the iceberg' of the complete analysis. The purpose of the presentation and the audience largely determine what analysis is presented…" (2012, p. 62). For example, law enforcement agencies routinely post information produced from ACA on their websites in the form of community bulletins, interactive web-based maps, or agency reports.

3.2.2. *Tactical crime analysis*. Tactical Crime Analysis (TCA) emphasizes collecting data, identifying patterns, and developing possible leads so that criminal cases can be cleared quickly. TCA usually involves analysis of individual, incident-level data associated with specific events (i.e., robberies, motor vehicle thefts, residential burglaries, etc.). Analysts engaged in TCA often produce reports containing time series or point-pattern information depicted in charts, graphs, maps, or a combination of each. In short, TCA is a crime analysis technique that aims to describe and convey information about crime patterns quickly and easily so that the effects of crime fighting and reduction strategies can be maximized.

3.2.3. *Strategic crime analysis.* Unlike the other two approaches, Strategic Crime Analysis (SCA) is focused on operational strategies in an attempt to develop solutions to *chronic* crime-related problems. Spatial analytic techniques associated with SCA usually involve analysis of geographic units (i.e., jurisdiction, census tract, patrol district, beat, etc.). SCA focuses on clusters analysis in order to produce information that can be used for resource allocation, beat configuration, the identification non-random patterns in criminal activity, and unusual community conditions. In short, SCA provides law enforcement agencies with the ability to provide more effective and efficient service to the community. One of the most popular analytic techniques used in SCA is "hot spot" analysis, which is described in the following section in greater detail.

3.3. Crime Hotspot Analysis

Areas of concentrated crime are often referred to as crime hotspots (e.g. McLafferty et al., 2000, Eck et al., 2005). Crime analysts look for concentrations of individual events that might indicate a series of related crimes; they also look at small areas that have a great deal of crime even though there may not be a common offender. The common understanding is that a hotspot is an area that has a greater than average number of criminal events, or an area where people have a higher than average risk of victimization (Chainey & Ratcliffe, 2005).

There are different levels of hotspots analysis (Eck et al., 2005), depending on the size of the geographic area of concern, from very specific locations or addresses, to blocks, street and neighborhoods. Each level corresponds to a particular question being addressed. Underlying the analysis of crime hotspots at these levels are several crime theories, which range from theories on the social ecology of crime to theories on routine activities and repeat victimization (Eck et al., 2005; Anselin et al., 2000).

A wide range of different methods and techniques has emerged to characterize crime hotspots and a solid review is provided by Eck et al. (2005). The techniques fall into three different categories: 1) global statistical tests, such as mean center, standard deviation distance and ellipse, and global tests for clustering, including the Nearest Neighbor Index, Moran's I and Geary's C statistic; 2) hotspots mapping techniques, such as point mapping, spatial ellipses using hierarchical or K-means clustering, thematic mapping using enumeration areas, quadrat mapping, and kernel density estimation; and 3) local indicators of spatial association statistics, such as the Gi and Gi* statistics.

While many of these techniques serve a somewhat different purpose, they are all concerned with characterizing hotspots in an effort to develop a better understanding of where crimes occur, which can ultimately lead to the design of intervention strategies and the development of prospective crime mapping. No single technique has emerged as the "best" one for crime hotspot mapping, and there has been surprisingly little comparative research on their strengths and weaknesses, with some notable exceptions (Chainey, 2005; Chainey et al., 2008).

Most crime hotspot analysis techniques are based on a dataset of individual locations, with each point representing one or multiple crime incidents. The dataset of crime incidents is assumed to be a very good representation of the actual crimes incident (i.e., the sample is complete or very close to complete and the locations are accurate). These locations are usually derived through the geocoding of the address information in the crime incident reports. While the comparison of different hotspot analysis techniques has received some attention in the literature (e.g., Grubesic, 2006), very little attention has been paid to the quality of the geocoding process and its effect on hotspot analysis.

One of the unique characteristics of crime incidents is that they occur in many different types of locations, including private residences, office buildings, public places, along the road network, etc. This presents unique challenges both for the geocoding of crime incidents and their analysis.

For example, any traffic-related incidents are logically located along the street network, and this presents some unique challenges (e.g., Levine & Kim, 1999). Geocoding descriptive information, such as “150 meters south of the intersection of Main Street and 4th Ave” is cumbersome using automated methods and is therefore often accomplished by geocoding the nearest intersection instead. Spatial analysis techniques used to identify patterns in traffic related incidents will need to distinguish the clustering of events from the clustering of the street network itself. Hotspot detection techniques for networks have received some attention (e.g., Anderson, 2006; Tompson, Partridge, & Shepherd, 2009), but have been widely used or tested.

A second characteristic of crime incidents is that they often have a tendency to occur at the exact same location (repeat burglaries, incidents at intersections, etc.). This presents both opportunities for spatial crime analysis (e.g., Ratcliffe & McCullagh (1998) used GIS to identify repeat victimizing) as well as computation challenges to cluster detection (Brimicombe, 2005).

3.4. Predictive Crime Mapping

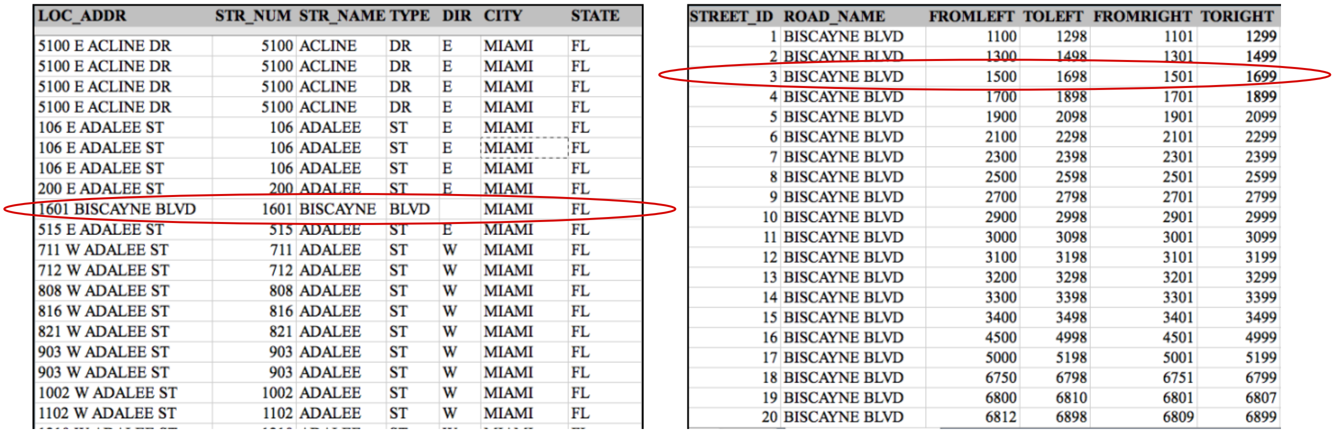
While most applications of crime mapping are retrospective, there is a growing interest in using crime mapping techniques for predicting future crime events in space and time, with the ultimate goal of informing a proactive approach to crime prevention. While the field of predictive crime mapping is relatively new, the interest is clearly growing and there have been a number of recent publications (e.g., Bowers et al., 2004; Chainey et al., 2008; Christens & Speer, 2005; Groff & LaVigne, 2002; Johnson & Bowers, 2004; Johnson et al., 2009). The reliability of these techniques, however, also depends heavily on the accuracy of the retrospective hotspot maps created from past events.

Several general approaches exist in predictive crime mapping, including the use of temporally aggregated hotspots, individual-level analysis of repeat victimization, and various univariate and multivariate analysis of area level data – see Groff and LaVigne (2000) for a review. While there is no agreement on which general approach is most reliable for crime prediction, the research appears to suggest that the “best” method is likely to depend on the type of crime. For example, Johnson et al. (2009) successfully developed a predictive individual-level model based on optimal foraging behavior of repeat offenders, but the approach was specifically designed for burglaries and may not apply to other types of offenses. Similarly, Johnson and colleagues (2008) determined substantial variation in the stability of crime hotspots and this is expected to vary by type of crime.

Among the various approaches to predictive crime mapping, hotspot analysis has received the most attention—in part because many hotspot techniques are in widespread use, in part because of their versatility across spatial-temporal scales and across types of crime. Chainey et al. (2008a) provided one of the first comparative analyses of a range of hotspots techniques for predictive crime mapping and introduced the concept of a Predictive Accuracy Index (PAI). This provides a measure of how reliable a retrospective hotspot is able to predict future crime events relative to the size of the hotspots. In a response by Levine (2008) this was extended with the use of the Recapture Rate Index (RRI). These two indices provide a solid foundation for a more comprehensive comparison of predictive hotspots methods across study areas—research being undertaken by NIJ (Wilson, 2009). What is still missing, however, is a consideration of data quality—hotspots are created from geocoded crime events, and both incomplete geocoding and positional error may have substantial effects of the robustness of hotspot methods.

3.5. Background on Geocoding

Addresses are one of the fundamental means by which people conceptualize location in the modern world. In a Geographic Information System (GIS) addresses are converted to features on a map through the geocoding process. Geocoding is the process of assigning an XY coordinate pair to the description of a place by comparing the descriptive location-specific elements to those in reference data. Figure 1 illustrates this process.



LOC_ADDR	STR_NUM	STR_NAME	TYPE	DIR	CITY	STATE
5100 E ACLINE DR	5100	ACLINE	DR	E	MIAMI	FL
5100 E ACLINE DR	5100	ACLINE	DR	E	MIAMI	FL
5100 E ACLINE DR	5100	ACLINE	DR	E	MIAMI	FL
5100 E ACLINE DR	5100	ACLINE	DR	E	MIAMI	FL
106 E ADALEE ST	106	ADALEE	ST	E	MIAMI	FL
106 E ADALEE ST	106	ADALEE	ST	E	MIAMI	FL
106 E ADALEE ST	106	ADALEE	ST	E	MIAMI	FL
200 E ADALEE ST	200	ADALEE	ST	E	MIAMI	FL
1601 BISCAYNE BLVD	1601	BISCAYNE	BLVD		MIAMI	FL
515 E ADALEE ST	515	ADALEE	ST	E	MIAMI	FL
711 W ADALEE ST	711	ADALEE	ST	W	MIAMI	FL
712 W ADALEE ST	712	ADALEE	ST	W	MIAMI	FL
808 W ADALEE ST	808	ADALEE	ST	W	MIAMI	FL
816 W ADALEE ST	816	ADALEE	ST	W	MIAMI	FL
821 W ADALEE ST	821	ADALEE	ST	W	MIAMI	FL
903 W ADALEE ST	903	ADALEE	ST	W	MIAMI	FL
903 W ADALEE ST	903	ADALEE	ST	W	MIAMI	FL
1002 W ADALEE ST	1002	ADALEE	ST	W	MIAMI	FL
1102 W ADALEE ST	1102	ADALEE	ST	W	MIAMI	FL

STREET_ID	ROAD_NAME	FROMLEFT	TOLEFT	FROMRIGHT	TORIGHT
1	BISCAYNE BLVD	1100	1298	1101	1299
2	BISCAYNE BLVD	1300	1498	1301	1499
3	BISCAYNE BLVD	1500	1698	1501	1699
4	BISCAYNE BLVD	1700	1898	1701	1899
5	BISCAYNE BLVD	1900	2098	1901	2099
6	BISCAYNE BLVD	2100	2298	2101	2299
7	BISCAYNE BLVD	2300	2398	2301	2399
8	BISCAYNE BLVD	2500	2598	2501	2599
9	BISCAYNE BLVD	2700	2798	2701	2799
10	BISCAYNE BLVD	2900	2998	2901	2999
11	BISCAYNE BLVD	3000	3098	3001	3099
12	BISCAYNE BLVD	3100	3198	3101	3199
13	BISCAYNE BLVD	3200	3298	3201	3299
14	BISCAYNE BLVD	3300	3398	3301	3399
15	BISCAYNE BLVD	3400	3498	3401	3499
16	BISCAYNE BLVD	4500	4998	4501	4999
17	BISCAYNE BLVD	5000	5198	5001	5199
18	BISCAYNE BLVD	6750	6798	6751	6799
19	BISCAYNE BLVD	6800	6810	6801	6807
20	BISCAYNE BLVD	6812	6898	6809	6899

Figure 1. Illustration of the geocoding process, whereby a latitude and longitude (i.e., XY coordinate) is assigned to an input address (left), based on the known location of reference address information (right).

The geocoding process is defined as the steps involved in translating an address entry, searching for the address in the reference data, and delivering the best candidate or candidates as a point feature on the map.[2] Specifically, the geocoding process involves:

1. Parsing the input address into individual address elements based on an address locator service.[3] Figure 2 provides an example of the different elements of an address that are used by a locator during the geocoding process, in order to establish a match candidate.

2. Standardizing address elements into abbreviations (i.e., “Street” into “St” or “Drive”

[2] Although different software applications have unique routines for executing a geocoding process, they all follow the same general steps listed in items 1- 9.

[3] An address locator service is a file containing style-specific guidelines and location-specific reference data, and is used within a GIS to interpret address input information in order to assign an XY coordinate pair. The current study used several locators, constructed from both reference data obtained from local jurisdictions as well as the data provide by the commercial vendors.

into "Dr").

3. Assigning address elements to match keys or particular categories that are used to compare to the categories contained in the reference data. For example, in the address 1600 Pennsylvania Ave., "1600" would be assigned to the "house number" category and "Ave" would be assigned to the "street type" category.

4. Calculating index values for some elements of the address, which are used to compare against a geocoding index. Indexing helps to increase processing speed.

5. Searching the reference data for features that contain similar elements to those in the input address.

6. Scoring potential matches identified in the reference data.

7. Listing candidates based on a user-defined minimum match score.

8. Producing the user-defined output for the address identified as the best candidate (i.e., the candidate with the highest match score).

9. Producing a feature class for the address that permits other geoprocessing tasks.

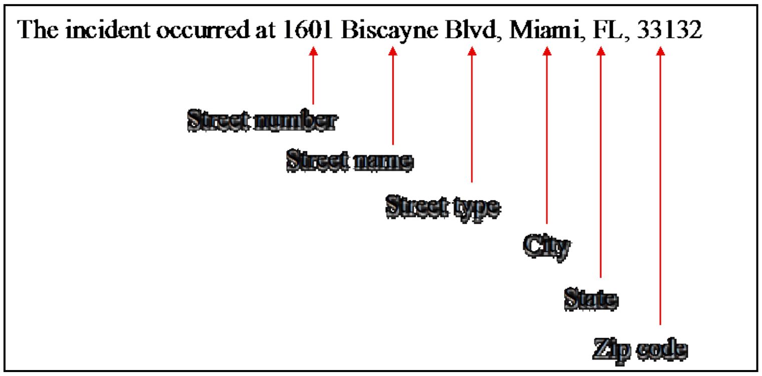

The incident occurred at 1601 Biscayne Blvd, Miami, FL, 33132
Street number
Street name
Street type
City
State
Zip code

Figure 2. Elements of an address used in the geocoding process.

Techniques involved in geocoding borrow from various academic fields, most notably, information theory, decision theory, probability theory, and phonetics. While geocoding applications are diverse and span many types of applications, there are several common problems associated with geocoding that have traditionally caused poor match rates, requiring excessive manual mapping by the user and potential inaccuracies and/or incompleteness in the resulting spatial datasets (e.g., Rushton et al., 2006; Goldberg et al., 2007)

One of the main challenges to accurate geocoding is the availability of good reference data. This requires a sturdy address model to organize the reference data components in a logical, maintainable and site-specific way. Several common address models exist. Each has a particular

set of supporting materials and characteristic errors.[4]
The first one can be characterized as the “geographic unit” model. These geographic units can consist of postal codes (such as ZIP codes in the US), counties, cities, census enumeration areas or any other geographic boundary considered meaningful. In the geocoding process, the location assigned to a particular address is the polygon (or the polygon centroid) representing the geographic unit.

The utility of the results is obviously related to the size of the geographic units. For example, in the United States 5-digit ZIP codes tend to be quite large, typically larger than census tracts, making them less attractive when spatially detailed information is required. When geocoding at the level of geographic units is not sufficient, several alternatives exist, including street networks, parcels, and address points. Each of these three address models will be described in more detail below.

The most widely employed address data model is based on a street network represented as street line segments that hold street names and the range of house numbers and block numbers on each side of the street. Address geocoding is accomplished by first matching the street name, then the segment that contains the house numbers and finally placing a point along the segment based on a linear interpolation within the range of house numbers. This approach to geocoding an address is referred to as “street geocoding” and has become the most widely used form of geocoding. Nearly all commercial firms providing geocoding services and most GIS software with geocoding capabilities rely primarily on street geocoding. Figure 3 provides a conceptual diagram of this process.

Parcels are traditionally the most spatially accurate data with address information available. Geocoding against parcels allows one to match against individual plots of land (or rather, the centroids of those polygons) rather than interpolating against a street centerline. This is particularly useful in areas where parcels are not regularly addressed (such as on roads with mixed parity) or those parcels that may be quite a distance from the centerline. Parcel geocoding typically results in a lower match rate in part because a single parcel can be associated with many addresses (Zandbergen, 2008a). Despite these lower match rates, parcel geocoding is considered more spatially accurate and is now becoming widespread given the development of parcel level databases by many cities and counties in the US (Rushton et al., 2006).

To overcome the limitations of parcels for geocoding, address points have emerged as a third address data model. Address points are commonly created from parcel centroids for all occupied parcels (or points can be placed elsewhere within the parcel, such as the location of the main structure or in front of the main structure). This is supplemented with address points for sub-addresses such as individual apartment units, condominium units, duplexes, etc. Field data collection or verification of building locations using digital aerial imagery can be used to further

[4] It should be noted that regardless of address model, a geocoded address represents a positional estimation of a “true” or actual location. As with any estimate, some degree of error will be associated with any geocoded address location. The goal, therefore, is to maximize the precision and accuracy of geocoding so as to improve spatio-temporal analysis of these data.

supplement the address point file. Address point data sets are of great value to local government, in particular emergency services. Figure 4 shows an example of an address point data file in a GIS environment, super-imposed on aerial imagery and parcel boundaries.

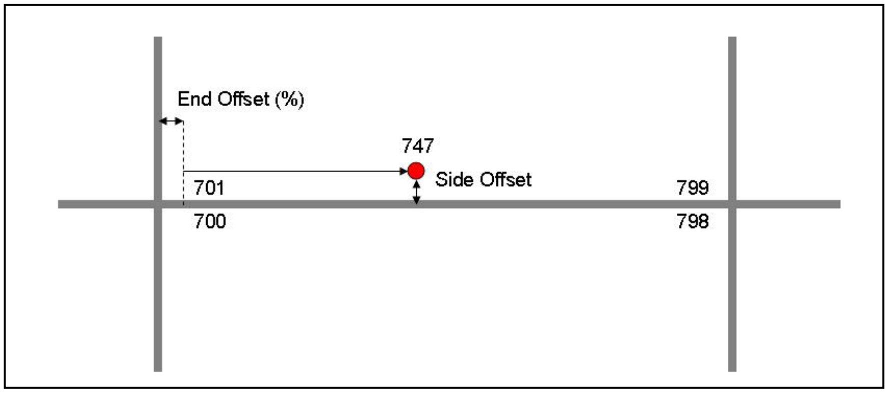


Figure 3. Conceptual diagram of the algorithm behind street geocoding. The location of an address is placed on a street segment based on linear interpolation along the street segment within the street number range for the segment. Optional considerations are the use of a side offset to place the location at one side of the street and an end offset to "squeeze" the locations away from the end of the street segment (avoiding the placement of locations at intersections).

Although countries like Australia, Canada and the United Kingdom have already developed national address point databases, in the United States, address point geocoding is not in very widespread use at present. However, many local governments have started to create address point databases and several commercial geocoding firms provide address point geocoding for

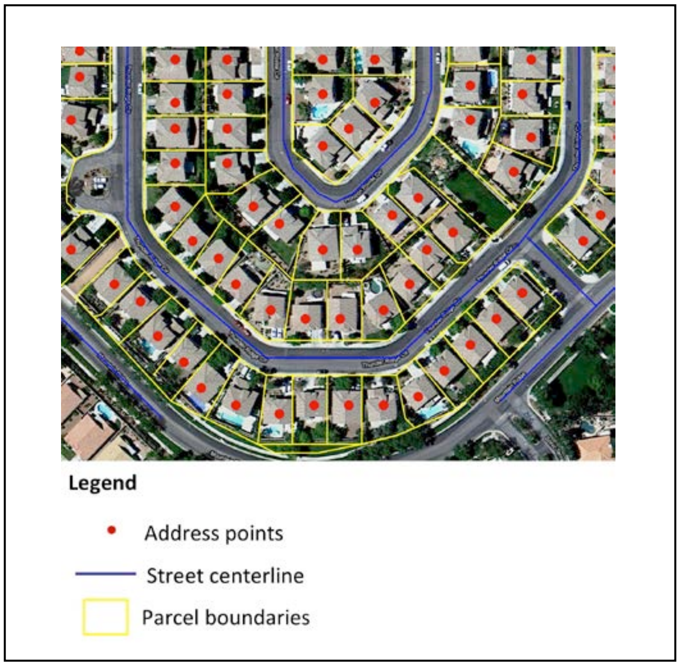


Figure 4. Example of address points and parcel boundaries for single-family residential area in Henderson, Nevada. Address points are typically placed either at the center of the residential structure or directly in front of it. The example only shows single-family residential housing; for other housing types multiple address points can be placed within a single parcel.

selected urban areas. Commercial firms already claim that around 51 million address points are available for the US, covering a selected number of metropolitan areas. An evaluation of address points has demonstrated that match rates are very similar to those obtained by street geocoding while the positional accuracy is far superior (Zandbergen, 2008a).

Within crime mapping, geocoding to geographic units (postal codes, census enumeration units) (e.g., Bernasco & Nieuwbeerta, 2005; Britt et al., 2005, Cahill & Mulligan; LaGrange, 1999; Poulsen & Kennedy, 2004) and street geocoding (e.g., Andresen, 2006; Bichler & Balchak, 2007; Doran & Lees, 2005; Grubesic, 2006; Harada & Shimada, 2006) are most common. Very few crime-mapping efforts have employed parcel and address point geocoding, with some notable exceptions in recent literature (Brimicombe et al., 2007; Grubesic et al., 2007; Zandbergen & Hart, 2006; Zandbergen, 2008a).

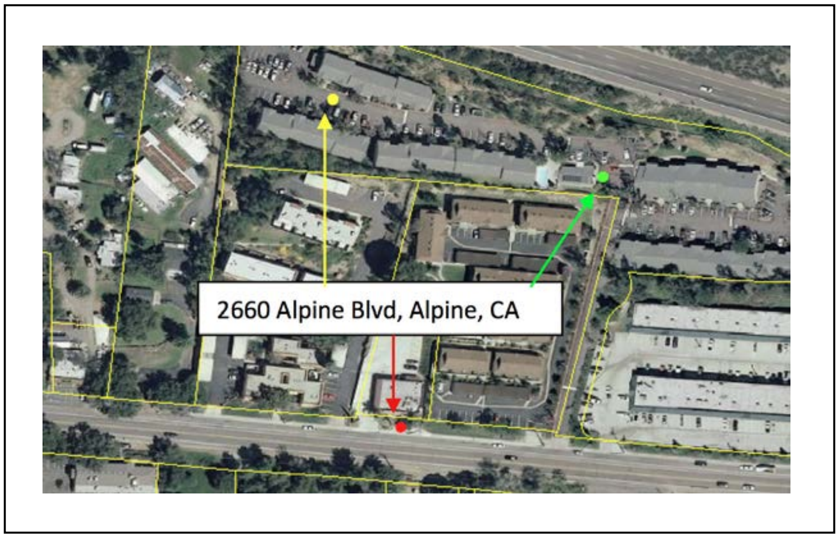


Figure 5. Illustration of an input address geocoded to three different reference layers. The red dot represents an address location geocoded against a street centerline file. The green dot represents the same address geocoded to a parcel centroid file, and the yellow dot represents the same address geocoded against address point reference data.

3.6. Geocoding Quality

Certain quality expectations must be met in order for the results of a geocoding process to be considered meaningful. That is, the overall quality of any geocoding result can be characterized in three distinct ways: completeness, positional accuracy, and repeatability. Completeness refers to the percentage of records that can reliably be geocoded, and is also commonly referred to as the match rate. Positional accuracy refers to how close each geocoded point is to the actual location of the address it is intended to represent (Figure 5). And finally, repeatability indicates how sensitive geocoding results are to variations in the input address, reference data, matching algorithms of the geocoding software, and the skills and interpretation of the analyst. In sum, geocoding results that are of high quality are complete, spatially accurate, and repeatable.

The match rate or the proportion of addresses reliably geocoded relative to all addresses available for geocoding represents the simplest measure of geocoding quality. Many factors can influence match rates and studies that have employed geocoding report match rates that vary considerably. Although some criminal justice research attempts to establish a minimally acceptable geocoding match rates (Ratcliffe, 2004), there is no consensus on a universal standard for this figure.

The lack of an established match rate threshold is due in part to the fact that interpreting match rates is very subjective, since much depends on the criteria used to characterize a “match”. For example, a higher match rate can easily be accomplished by simply lowering the minimum match score required to produce a match. However, lowing the minimum match score may inadvertently introduce false positives into geocoding results. Simply put, for any given set of input addresses used in the geocoding process, there is a trade-off: increasing the match rate by lowering the minimum match score results in a decrease in accuracy and therefore the overall quality of geocoded locations.

One additional dimension to geocoding completeness is the potential bias introduced by an incomplete result. It is well established, for example, that match rates are lower in rural areas (e.g., Cayo & Talbot, 2004; Zandbergen, 2011) in part due to the use of postal routes instead of street addresses in rural areas. These differences in match rates across urban/rural gradients can lead to substantial bias in spatial analysis (e.g. Oliver, Matthew, Siadaty, Hauck, & Pickle, 2005). Other factors affecting match rates have not received as much attention, but at least one study (Gilboa, Mendola, Olshan, Harness, Loomis, Langlois, Savitz, & Herring, 2006) found evidence of selection bias in terms of ethnicity that was produced from incomplete geocoded records.

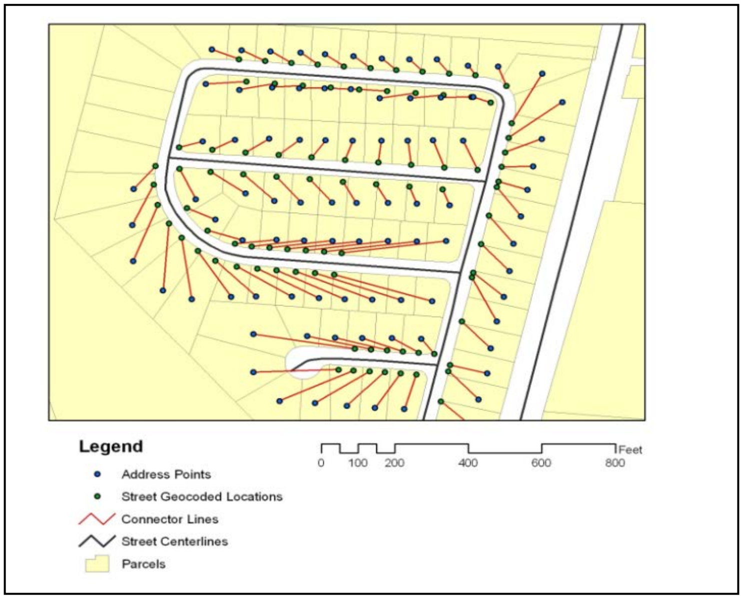


Figure 6. Positional accuracy illustrated. The blue dots represent a known address point, the green dot represents a street geocoded location of a crime incident, and the red line represents the positional error measured in Euclidean

distance between the two locations.

Several studies have determined quantitative estimates of the positional accuracy of geocoding. In a review of 12 different investigations, Zandbergen (2009) found that estimates of ‘typical’ positional errors for residential addresses ranged from 25-168 meters (Bonner, Han, Nie, Rogerson, Vena, & Freudenheim, 2003; Cayo & Talbot 2003; Dearwent, Jacobs, & Halbert, 2001; Karimi & Durcik 2004; Ratcliffe, 2001; Schootman, Sterling, Struthers, yan, Laboube, Emo, & Higgs, 2007; Strickland, Siffel, Gardner, Berzen, & Correa, 2007; Ward, Nucklos, Giglierano, Bonner, Wolter, Airola, Mix, Colt, & Hartge, 2005; Whitsel, Rose, Wood, Henley, Liao, & Heiss, 2006; Zandbergen, 2007; Zhan, Brender, De Lima, Suarez, & Langlois, 2006; Zimmerman, Fang, Mazumdar, & Rushton, 2007) based on median values of the error distribution. In addition, research suggests that results in urban areas are generally more accurate than in rural areas (Bonner et al. 2003; Cayo & Talbot 2003; Ward et al. 2005) and that the occurrence of major positional errors is relatively common. For example, in one of the more thorough studies by Cayo and Talbot (2003), 10% of a sample of urban addresses geocoded with errors larger than approximately 96 meters and 5% geocoded with errors larger than 152 meters. For rural addresses, these distances were 1.5 and 2.9 kilometers, respectively.

Relative to match rates and positional accuracy, geocoding quality described in terms of repeatability has not received as much attention from the scientific community. Nevertheless, in a recent study by Whitsel et al. (2006), substantial differences in results were identified when a large sample (n=3,615) of addresses from 49 U.S. States were geocoded across four different commercial vendors. Specifically, significant differences in address match rates (30%-90%), concordance between established and vendor-assigned census tracts (85%-98%), and distance between established and vendor assigned coordinates (mean of 228-1,809 meters) were identified. Conversely, in a comparison of three geocoding algorithms (LocMatch, ArcView 3.2, and Tele Atlas North America) using the same TIGER reference data, Kairimi and Durcik (2004) found that the differences between the results were not significant. This suggests that differences in reference data are at least in part responsible for the observed differences between commercial vendors. Utilizing three different street reference datasets, Zandbergen (2011) examined one aspect of repeatability and found that match rates and positional accuracy were highest for local street centerlines than for other types of non-commercial reference data. Finally, only one known study using crime data has examined geocoding quality in terms of repeatability. Specifically, Bichler and Balchak (2007) documented several specific limitations to the repeatability of geocoding, including the accuracy of reference database, choice of GIS software, and user-selected settings in the geocoding process.

3.7. Effects of Geocoding Quality on Spatial Analysis

Collectively, geocoding quality research clearly demonstrates that errors in geocoding can be very substantial and needs to be characterized in a meaningful manner relevant to the use of the geocoding results. Zandbergen (2009) provides a more thorough review of the effects of geocoding quality on spatial analysis, but a brief summary follows.

Errors in geocoded addresses may adversely affect spatio-temporal analyses, but this has not received widespread attention in the literature. Inflation of standard errors of parameters estimates as well as a reduction in the power to detect such spatial features as clusters and trends are among the specific effects that geocoding errors may produce (Jacquez & Waller, 2000; Waller, 1996; Zimmeman, 2007). Burra, Jerrett, Burnett, and Anderson (2002) demonstrated that even relatively small positional errors can have an impact on local statistics for detecting clusters. However, research on this topic has been mostly confined to the health field. For example, typical street geocoding is not sufficiently accurate for the analysis of exposure to traffic-related air pollution of children at short distances of 250-500 meters (Zandbergen, 2007; Zandbergen & Green, 2007). Similar errors in misclassification of exposure potential have been identified by Whitsel et al. (2006).

A growing number of studies in the crime literature have tried to determine the effect of geocoding quality on the results of crime analysis. Ratcliffe (2004) examined the effect of geocoding match rate on the resulting pattern in crimes rates aggregated to census boundaries and determined 85% as the minimum acceptable match rate. However, this analysis assumed there is no bias in the pattern of the ungeocoded locations. Brimicombe et al. (2007) demonstrated for a large metropolitan area in the United Kingdom that different geocoding match rates for the same crime incident database revealed distinct kernel density hotspots, although no statistical comparisons were made. Harada and Shimada (2006) compared kernel density surfaces derived from geocoded crime locations of different positional accuracy. Hotspots appeared relatively robust, although this can partially be attributed to the large bandwidth used (500 meters). Zandbergen and Hart (2009b) demonstrated that traditional street geocoding is insufficiently accurate to determine residency restrictions for sex offenders, due to the large number of false positives and negatives introduced. And finally, Zandbergen, Hart, Lenzer, and Camponovo (2012) recently showed how the misplacement of street geocoded crime incidents adversely impacts kernel density hotspot mapping techniques.

To summarize, the growing interest and use of crime mapping and analysis among practitioners and academics alike, the potential impact that geocoding results can have on the spatio-temporal analysis of crime, and a sparse literature in this area has given rise to the current study, which was divided into two complementary components. Recall, the aim of the first component of our research was to investigate the relationship between reference data and geocoding quality for crime analysis, whereas the second component of our research was designed to determine the impact of data quality on predictive hotspot mapping techniques. This next of the report presents findings from the geocoding quality analysis, followed by a section containing the results of our predictive hotspot analysis.

4. GEOCODING QUALITY ANALYSIS

4.1. Research Questions

We began our investigation with an analysis of reference data on geocoding quality. In particular, the completeness and positional accuracy of geocoding results was assessed across several crime types as well as different levels of input address quality, using both commercial and non-commercial street network reference data.[5] Analysis was preformed while controlling for specific user-defined parameters utilized by software during the geocoding process[6] in order to answer the following four research questions:

1. Does the completeness of geocoded crime data (i.e., the match rate) vary by crime type and/or input address quality?

2. Is the completeness of geocoded crime data influenced by the type of street reference data utilized in the geocoding process (i.e., commercial versus non-commercial)?

3. Does the positional accuracy of street geocoding vary by crime type and/or input address quality?

4. Is the positional accuracy of street geocoding influenced by the type of street reference data?[7]

Results from our geocoding quality analysis provide researchers and practitioners with valuable guidance and insight into one of the most basic—albeit fundamental—procedure related to the spatio-temporal analysis of crime, and provide direction for future research in this area. The data and methods employed in our research are described in the next section.

4.2. Data and Methodology

In order to answer the research questions presented in Section 4.1, we analyzed existing data from six large law enforcement jurisdictions in the U.S.: Arlington Police Department (TX), Albuquerque Police Department (NM), Charlotte-Mecklenburg Police Department (NC), Las Vegas Metropolitan Police Department (NV), San Diego County Sheriff’s Office (CA), and Tampa Police Department (FL).

[5] Variations in the input address, matching algorithms of the geocoding software, and the skills and interpretation of the analyst represent other aspects of geocoding that can affect the quality of geocoded results; however, the current study was limited to examining the affects that different types of reference data had on geocoding results.

[6] The number of parameters that can be manipulated combined with the breadth of settings to which these parameters can be set to is exponentially large. Although the affects of user-defined parameter settings are not examined in the current study, we feel that this particular aspect of geocoding quality as it relates to geocoding crime data warrants further investigation.

[7] Results for each of four research questions are provided for each jurisdiction in Appendix A.

These jurisdictions were chosen for several reasons, including the availability of existing data, the extent to which GIS is currently being used within these agencies, the size of the agencies, and the depth and breath of information associated with incident-level data. Since these agencies were not selected randomly, findings from the current study are representative of all jurisdictions. Nevertheless, three unique types of existing datasets were used in our analyses, including crime incident information for events that were recorded by each jurisdiction between 2007 and 2008, street network reference data covering each jurisdictional boundary, and address point and/or parcel reference data associated with properties located within each jurisdiction.

Data from all six jurisdictions were combined[8] to represent approximately 400,000 crime incidents known to law enforcement, obtained directly from the agencies listed previously (see Table 1). Data files contained information on the location of each crime event (i.e., the input address), the date and time that the incident occurred, the type of incident (i.e., crime type), a case/incident number, and in some cases the UCR/NIBRS classification code, incident status, as well as the patrol division, beat, or sub-beat in which the event occurred. In order to ensure consistency across jurisdictions' data sets, crime types were conceptualized using Part I crime definitions from the FBI's UCR Program.[9]

Data were provided in Excel format and restricted to only those incidents known to law enforcement and recorded as an aggravated or simple assault (n=54,592), auto burglary (n=102,213), auto theft (n=65,112), burglary (n=95,675), drug offenses (n=55,587), homicide (n=620), or robbery (n=24,294). In accordance with FBI reporting standards, only the most severe crime was included in the analysis for events involving more than one crime type. These data have been archived at the ICPSR (http://www.icpsr.umich.edu).

A total of five different street network layers were utilized as reference data for geocoding in the current analysis, including two commercial data files and three non-commercial files. The two data files used to support the commercial street geocoding processes included ESRI's StreetMap™ Premium based on Tele Atlas (2010 Release 1) data and ESRI's StreetMap™ Premium based on NAVTEQ (2010 Release 1) data. Non-commercial street network data used in the current study included ESRI's StreetMap™ USA files[10], Census 2009 TIGER/Line® files, and street centerline data provided by local government agencies.

[8] Match rate results for each agency are presented in Appendix A and results of the positional accuracy analysis for each agency are presented in Appendix B.

[9] Simple assault, auto burglary, and drug offenses are included in the current analysis, but are not Part I crimes. These incidents may vary somewhat by jurisdiction based on each agency's recording practices and information contained in their respective Records Management System (RMS). Simple assault includes a physical attack by an unarmed offender and that does not result in injury; auto burglary includes thefts of property of any value from a secured vehicle; and drug offenses include drug possession, trafficking, and narcotics-related incidents known to law enforcement.

[10] Although these reference data are not free to everyone, they are included with the purchase of ArcGIS software at no additional cost and represent older versions of commercial reference data such as Tele Atlas enhanced with TIGER/Line® data. For the purposes of the current study, these data are considered "non-commercial" street network data.

Table 1. Crime incidents by crime type and jurisdiction, 2007-08.

		Jurisdictions					
				Charlotte-		San Diego	
Crime events in –	Total	Albuquerque	Arlington	Mecklenburg	Las Vegas	County	Tampa
All jurisdictions							
All locations	391,997	57,172	24,749	103,064	120,779	40,880	45,353
Without intersections	354,587	56,646	24,726	97,702	103,344	37,173	34,996
Assaults	45,350	8,146	1,116	18,113	2,914	3,535	11,526
Auto burglary	95,454	17,676	13,158	29,982	19,354	9,528	5,756
Auto theft	62,482	11,442	2,752	12,121	26,248	6,560	3,359
Burglary	93,606	12,987	6,065	23,671	35,242	8,106	7,535
Drug offenses	35,957	3,134	426	7,782	11,619	8,170	4,826
Homicide	546	86	36	128	223	28	45
Robbery	21,192	3,175	1,173	5,905	7,744	1,246	1,949
Intersections only	37,410	526	23	5,362	17,435	3,707	10,357
Assaults	6,702	173	1	3,729	630	674	1,495
Auto burglary	3,203	121	11	182	1,729	389	771
Auto theft	2,630	83	5	197	1,682	281	382
Burglary	2,069	23	0	911	344	77	714
Drug offenses	19,630	64	1	15	10,774	2,110	6,666
Homicide	74	0	0	23	36	5	10
Robbery	3,102	62	5	305	2,240	171	319

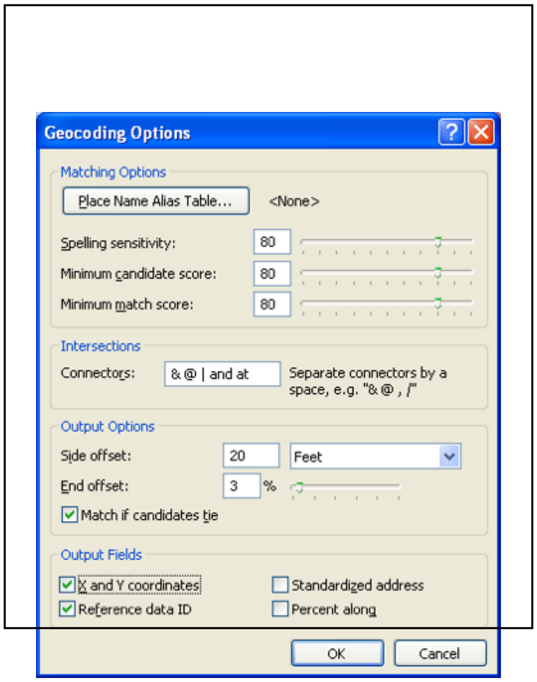
Note: Drug offenses include drug possession, trafficking, and narcotics calls for service.

To address the research questions related to geocoding completeness, only the crime event data and street centerline layers were used in the analyses.[11] Consideration was given to variations in the quality of street centerline files by street geocoding each crime incident against each of the five street data layers noted previously. Geocoding procedures were held constant across all jurisdictions and included creating an address locator service for each type of reference data. A U.S. Address—Dual Ranges locator style was constructed for the street centerline reference data and a U.S. Address—Single House locator style was used for the parcel/address point reference files. All address fields contained in each locator style were constructed using the same input fields and other user-defined parameters considered during the geocoding process (i.e., spelling sensitivity (80), minimum candidate score (80), street offset (20 feet), end offset (3%), intersection connectors, etc.) were held constant across all procedures. Figure 6 provides a screen capture of the user-defined settings that were held constant during the geocoding process.

A crime incident location was considered "matched" if a matched or tied street geocoded location was interpolated from the street centerline layer with a match score of at least 80%. Overall "match rates" were calculated by dividing all matched and tied street geocoded crime locations by all crime records contained in the data file. The result of this process was a match rate for each crime type (i.e., homicide, robbery, drug offenses, etc.), determined for each street centerline reference layer (i.e., Tele Atlas, NAVTEQ, TIGER/Line®, StreetMap™ USA, etc.), broken down by two categories of input addresses: "Without intersections" and "Intersections

[11] All geocoded data used in the current study were geocoded using the same software (ArcGIS 10.0). Using the same software assured that results were not due to variations in the geocoding algorithms.

only”.[12] The “Without intersections” input addresses exclude intersections and addresses where the house number is represented as a 100-block number (i.e., 1600 Block of Pennsylvania Ave.), whereas the “Intersections only” input addresses include only these types of locations.



Geocoding Options

Matching Options

<None>

Spelling sensitivity: 80

Minimum candidate score: 80

Minimum match score: 80

Intersections

Connectors: & @ | and at

Separate connectors by a space, e.g. "& @ , /"

Output Options

Side offset: 20 Feet

End offset: 3 %

☑ Match if candidates tie

Output Fields

☑ X and Y coordinates

☑ Reference data ID

☐ Standardized address

☐ Percent along

Figure 7. Example of a Geocoding Options interface where a user can manipulate settings such as spelling sensitivity, minimum candidate scores, minimum match scores, offsets, etc. These settings can dramatically affect geocoding results and were therefore held constant across all geocoding procedures in the current study.

Address point reference data were obtained from local authorities responsible for overseeing GIS needs for each jurisdiction included in the study (e.g., the Mecklenburg County Geospatial Information Services). Address points represent the locations of all addressable structure in a jurisdiction and are commonly placed directly on top of the specific building or directly in front of it. Large structures with multiple units (e.g., shopping malls and apartment complexes) often have multiple address points to represent individual units. When address point information was not available for reference, a parcel reference data layer was employed. Like address point information, parcel layers were obtained from local government agencies. For the analysis of positional error, crime incidents geocoded against either the address point or parcel reference data represented the best estimate of where the crime event took place.[13]

[12] Most incidents are associated with crimes that occurred in or around an addressable structure (92%). Therefore, it is likely that a reliable address was reported for those events. Some crimes, however, occur on the street, in open spaces, or between structures. In these cases, address information may be less reliable. We expect that the balance between these two types of events will be different for different crime types, which would explain some variation in match rates between types of reference data. For these reasons, we have produced match rates across the two types of input addresses.

[13] As noted in Section 2.1 above, geocoding crime events using address point reference data was the preferred approach. However, when address point data was unavailable, parcel reference data was used. When parcel reference data is used, the resulting geocoded locations are placed at the centroid of the parcel polygon.

In order to determine the positional accuracy of each street geocoded crime incident, results of the street geocoded locations were compared to the corresponding address point/parcel geocoded location. Positional error for a specific event was determined as the Euclidean distance (in meters) between two points: 1) the street geocoded location and 2) the reference location based on the address point or parcel. This was carried out separately for each of the five different street geocoding techniques. The positional accuracy assessment employed only those crime events that were successfully geocoded across all street geocoding techniques as well as to the address point or parcel. Since comparisons between street and address point/parcel geocoded locations cannot be made for locations classified as intersections, incidents described as intersections only were excluded from the positional accuracy analysis[14].

Traditionally, different statistical estimates are used to characterize positional accuracy, including the range, mean, median, standard deviation, Root Mean Square Error (RMSE), and percentiles (75th, 90th, 95th, etc.). Although there is no agreement in the published research on the most meaningful statistic, several studies that report the actual distribution of error estimates (Cayo & Talbot, 2003; Karimi & Durcik, 2004; Zandbergen, 2008b) suggest that this distribution is log-normally distributed. This indicates that traditional statistics such as range, mean, standard deviation, and RMSE are not very meaningful. Therefore, in the current study, the positional accuracy of the street geocoded crime incidents are described in terms of median values (in meters) and the 95th percentile of positional error (in meters). The next section describes our results.

4.3. Results

As noted above, one way to determine the geocoding quality of crime event locations is by examining the completeness of the geocoded data. Completeness is the percentage of records that are reliably geocoded, also referred to as the match rate. Generally, geocoded data with higher match rates are considered by many to be more desirable than geocoded data with lower match rates. Therefore, the first stage of our geocoding quality analysis involved assessing the completeness of geocoded crime incidents in order to determine whether match rates vary by crime type and whether completeness of geocoded crime data is influenced by the quality of input address information and/or the type of street reference data utilized during the geocoding process. These findings provide answers to our first two research questions.

4.3.1. *Match rates*. Table 2 presents match rates for street geocoding results for each type of crime and each type of reference data considered. These results are further broken into two groups: input addresses that exclude intersections (i.e., “Without intersections”) and input addresses comprised of 100-block house numbers or intersections (i.e., “Intersections only”). Results are for all crime events that were recorded for 2007 and 2008, within each of the six law

[14] Although a point layer can be constructed from intersections contained in each of the street centerline files and could be used to represent the “true” location of “Intersection only” events, these nodes would represent both the “true” location of an incident and the street geocoded location for incidents geocoded to them. As a result, the positional accuracy of the intersection data would be perfect (i.e., no positional error) and would bias downward our aggregate measure of positional accuracy.

enforcement agency jurisdictions included in our study.

Findings show that street geocoding match rates vary by crime type and that this variation is influenced by the quality of address information being geocoded. When addresses expressed as an intersection or as a 100-block are excluded from the analysis, for example, match rates range from a low of 83% of auto burglaries, drug offenses, and homicides to a high of 86% of robberies. However, when only intersection data are considered, the match rate decreases dramatically and the variability of match rates across crime types increases. For example, among these locations, 15% of burglaries are successfully geocoded compared to 39% of robberies. These results demonstrate a direct correlation between the level of detail (i.e., an input address with a specific house number as opposed to a house number described as a block number or the address described as an intersection) associated with the input address and the geocoding match rates: when geocoding procedures are held constant, geocoding quality measured in terms of match rates generally increases when the quality of the input address is also increased. And although the same crimes are not consistently associated with the highest or the lowest match rates across each jurisdiction included in the study, the inverse relationship between input address quality and match rates is consistently observed across all locations (see Tables A1 through A6 in Appendix A). Regardless of jurisdiction, input addresses classified as "Intersections only" are successfully geocoded at considerably lower rates than addresses classified as "Without intersections." Interestingly, the magnitude of effect that input address quality has on geocoding results is not consistent across particular types of street reference data.

Table 2. Geocoding match rate results by crime type and type of street reference data, 2007-08.

		Street Geocoding				
		Free			Commercial	
		Local	StreetMap			
Crime events in --	Average	Centerlines	USA	TIGER	NAVTEQ	Tele Atlas
All jurisdictions						
All locations	78.8	86.4	72.3	69.9	82.6	82.7
Without intersections	84.4	92.9	77.1	74.4	89.1	88.7
Assaults	85.8	92.8	79.4	76.2	90.3	90.2
Auto burglary	83.2	92.8	75.0	72.5	88.1	87.6
Auto theft	84.9	92.3	77.5	74.6	89.8	89.9
Burglary	85.0	93.7	75.3	75.2	91.1	89.9
Drug offenses	82.8	92.0	81.6	74.6	83.0	82.9
Homicide	83.4	90.3	75.3	70.9	89.4	91.2
Robbery	85.9	93.0	81.4	74.1	89.7	91.4
Intersections only	25.3	24.8	26.3	27.7	21.7	25.8
Assaults	19.6	11.8	19.4	25.5	18.5	23.1
Auto burglary	25.9	27.2	27.9	28.9	18.0	27.6
Auto theft	37.4	39.4	38.1	39.7	29.8	39.8
Burglary	14.5	8.1	13.0	19.0	15.2	16.9
Drug offenses	23.8	25.7	25.4	24.9	20.4	22.5
Homicide	16.8	12.2	18.9	23.0	16.2	13.5
Robbery	43.2	44.1	43.9	44.3	38.6	45.3

Note: Drug offenses include drug possession, trafficking, and narcotics calls for service.

Greater disparities in geocoding quality—as measured by match rates—are observed between different types of street reference data, when it is considered in conjunction with different types of crimes and input addresses. For example, overall local centerline data and both commercial reference layers consistently have the highest match rates among all types of street reference data, regardless of the type of crime being geocoded. This pattern also holds true for every jurisdiction, with the exception of San Diego County (see Table A5 in Appendix A), where the commercial reference data underperformed more than any of the non-commercial data; and could suggest that findings from previous studies that have observed systematically lower match rates associated with more rural areas (e.g., Cayo & Talbot, 2004; Zandbergen, 2011) could be reference-data dependent.

When data are combined, for local centerline files, 86% of all addresses are successfully matched, compared to 83% of addresses matched against either commercial street file. Within the local centerline results in particular, when input addresses containing only intersection information are excluded, relatively higher match rates are observed. For example, the percent of addresses matched ranges from a high of 94% for burglary to a low of 90% for homicide. For the two commercial files, match rates are still higher relative to other types of reference data used, but the types of crimes associated with the highest and lowest match rates are different than the patterns observed for the street centerline reference data. Specifically, the highest match rates for the Tele Atlas data are associated with homicide (91%) and robbery (91%) and the lowest match rates are for drug offenses (83%). NAVTEQ data produce the highest match rate for burglary incidents (91%) and the lowest for drug offenses (83%). These results suggest that the quality of geocoding is influenced considerably by both the type of reference data used in the geocoding process as well as the particular type of crime incident being geocoded. Again, with one exception, this general pattern is also observed across each jurisdiction.

Finally, when input address information that contains only block numbers or intersections is considered independently of other addresses, geocoding quality in general decreases dramatically. Among these data, overall, none of the street centerline files used as reference data successfully geocodes more than 28% of the input addresses (TIGER/Line® 2009). However, greater variability in the match rate among the individual jurisdictions is observed. For example none of the “Intersections only” crime data successfully geocoded to the local street centerline file in Arlington, whereas 71% of these records matched against the local centerline and Tele Atlas reference data in Albuquerque (see Tables 2A and 1A, respectively).

Collectively, variation in match scores within each type of reference data is much more pronounced across specific crime types for the “Intersections only” data than for input addresses that exclude intersections. For example, among input addresses containing only block numbers or intersections, local street centerline reference data successfully geocodes as many as 44% of all robbery locations, but as few as 8% of all burglary locations. These findings show that when relatively lower quality input address information is street geocoded, crime type has a greater influence over match rates than does the type of reference data being used in the geocoding process.

4.3.2. *Positional accuracy*. A second measure of geocoding quality is positional accuracy, or how close a geocoded location is to the reference location of that event. Therefore, the next stage of our analysis involved assessing the positional accuracy of the street geocoded crime incidents for each reference data type in order to determine whether it varies by crime type and whether it is influenced by the type of street reference data utilized during the geocoding process. This stage of our analysis provides answers to our final two research questions.

Estimates of positional error are expressed in terms of two summary metrics of error distribution: median positional error (in meters) and the 95th percentile (in meters). Data used in the current analysis were produced from a subset of data (n=193,875) used in the match rate analysis and include each crime type: aggravated and simple assault (n=26,015), auto burglary (n=53,298), auto theft (n=34,296), burglary (n=51,725), drug offenses (n=17,265), homicide (n=284), and robbery (n=10,992). In order to produce an estimate of positional error for each type of reference data, a crime event was included in the current sample only if it was successfully street geocoded to each reference data layer as well as to a corresponding parcel or address point. Table 3 presents findings from the positional error analysis.

When all jurisdictions and all crime types are combined,[15] results show that the median error distance for all successfully street geocoded crime incidents ranges from a low of about 60 meters for the local centerline reference data to a high 81 meters for the StreetMap™ USA reference data, which is consistent with past research conducted with data from other disciplines (see Zandbergen, 2009 for a review). Between the two commercial reference products, the Tele Atlas data and the NAVTEQ data perform about the same: both have median positional error values for all crimes considered together around 65 meters. Overall, these levels of positional accuracy are consistent with distances observed for other types of address datasets, but demonstrate the variability across reference data types. And as with the match rate results, the general pattern observed for all the data combined is consistently observed for each jurisdiction: the StreetMap™ USA and the TIGER/Line® 2009 reference data consistently produce less accurate geocoding results than the other reference data considered (see Table B1 in Appendix

Table 3. Positional error statistics for street geocoded crime events in all jurisdictions, 2007-08.

		Free			Commercial	
Error/Crime type	Sample	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas
Median positional error (m)						
All crimes	193,875	60	81	70	64	65
Assaults	26,015	59	73	68	63	64
Auto burglary	53,298	62	88	76	69	70
Auto theft	34,296	63	85	72	66	67
Burglary	51,725	54	77	64	59	59
Drug offenses	17,265	67	78	72	65	67
Homicide	284	59	75	71	66	66
Robbery	10,992	70	84	77	67	70
95th percentile positional error (m)						
All crimes	193,875	263	322	298	244	412
Assaults	26,015	255	307	277	241	421
Auto burglary	53,298	274	349	330	257	434
Auto theft	34,296	283	355	312	254	434
Burglary	51,725	232	292	272	220	326

[15] Positional error estimates for each jurisdiction represented in the study are available in Appendix B.

B). Figure 6 presents the frequency distribution (in meters) for all crimes and jurisdictions combined for each type of reference data.

When geocoding quality is described for all the data combined in terms of the 95th percentile of positional error, overall, the two commercial products perform the best and the worst. For example, 5% of all crimes geocoded against the NAVTEQ data have positional errors greater than 244 meters, whereas 5% of all crimes geocoded against the Tele Atlas data have positional errors that exceed 412 meters. Among the free reference data, these errors range from 263 meters for local street centerline file to 322 meters for StreetMap™ USA data and are consistent with what has been observed in past research (Bonner et al., 2003; Cayo & Talbot, 2003; Ward et al., 2005; Zandbergen, 2011). With the exception of Albuquerque and San Diego County, the two commercial products also perform the best and the worst for each jurisdiction (see Tables A1 and A5, respectively).

Greater variability in positional accuracy across reference data is observed when different crime types are considered in conjunction with variations in reference data. Among specific crime types, the most accurate type of street geocoded locations is burglary incidents geocoded against local street centerline reference data (54 meters). On the other hand, the least accurate type of street geocoded crime data is auto burglaries geocoded against StreetMap™ USA reference layers (88 meters). And while there are a large number of small errors associated with each type of reference data, consistently 5% of street geocoded data have very large errors of several hundred meters[16]. For example, 5% of all crimes geocoded against Tele Atlas data are associated with error distances of greater than 1.1km for homicides, 5% of crimes geocoded against StreetMap™ USA data have error distances of greater than 355 meters for auto thefts, and 5% of crimes geocoded against TIGER/Line® 2009 data have error distances of greater than 330 meters for auto burglaries. This presents a persistent problem of incorrect locations, especially since the error distances are not random in nature (i.e., the error distribution is not normal). This means that substantially large error, in excess of several kilometers, is not uncommon. The final sections

[16] Detailed measures of positional accuracy for each crime type, including the minimum, maximum, median, 68th, 90th, 95th, and 99th percentiles of the error distributions are reported in Table C1 of Appendix C. This information covers the span of positional accuracy measures represented in the literature.

discuss the implication of these findings, identify specific limitations of the current study, offer suggestions for future research in this area, and provide recommendations to analysts on how to maximize the overall quality of street geocoded crime data.

4.4. Discussion and Conclusions

The current study provides answers to important questions related to the geoprocessing of crime event data. First, results of our analysis suggest that the type of reference data used in the geocoding process affects geocoding quality, measured in terms of completeness or match rates. For example, when all types of crime are considered together, findings show that local street centerline data outperforms all other street network reference data, including commercial data such as those produced by NAVTEQ and Tele Atlas.

Second, results of our analysis also suggest that match rates vary by crime type and are influenced greatly by the quality of input addresses. On average, for input addresses that exclude less detailed address information (i.e., 100-block house number addresses and addresses described as an intersection) match rates for robbery and assaults are higher than other crime types considered, whereas match rates for auto burglary, drug offenses, and homicide are somewhat lower. When only crime incident locations whose address information is less detailed (i.e., “intersections only”) are considered, overall match rates worsen dramatically; and, disparities in match rates by crime types become more pronounced. Among these results specifically, on average, findings show that robbery (43%) and auto theft (37%) is affected relatively less by the quality of input address than burglary (15%) and homicide (17%). And although the same crimes are not consistently associated with the highest or the lowest match rates across each jurisdiction included in the study, the inverse relationship between input address quality and match rates is consistently observed across all locations.

Third, findings from the current study suggest that the type of reference data in conjunction with crime type influences geocoding quality considerably; but consistent patterns across specific types of reference data are not apparent. For example, for input addresses that exclude intersections, local centerline and NAVTEQ reference data produce the best geocoding results for burglaries, StreetMap™ USA data produces the highest match rates for drug offenses, TIGER/Line® 2009 data geocodes assaults better than other crime types considered, and Tele Atlas data produces the best results for homicide and robbery. With few exceptions, the two commercial products also perform the best and the worst for each jurisdiction.

Fourth, results of our analysis suggest that the type of reference data used in the geocoding process affects geocoding quality, measured in terms of positional accuracy. For example, when all types of crime are considered together and positional accuracy is described in terms of median positional error, findings show that local street centerline data outperforms all other street network reference data, including commercial data such as those produced by NAVTEQ and Tele Atlas; and when positional accuracy is described in terms of the 95th percentile of the error distribution, local centerline data outperforms all other street network reference data, with the exception of NAVTEQ data.

Finally, findings from the current study also suggest that the positional accuracy of geocoded crime incidents is influenced greatly by crime type. For example, although local centerlines outperformed all other types of reference data considered when accuracy is measured in terms of median positional error, positional error ranges from a high of 70 meters for robbery to a low of 54 meters for burglary. Similar disparities are observed for the overall worse performing reference data. That is, the median positional error for all crimes geocoded against StreetMap™ USA data was 81 meters, but range from a low of 72 meters for auto theft to a high of 88 meters for auto burglary. Ironically, when positional error is measured in terms of the 95th percentile of the positional error distribution, the commercial reference data produce the most and the least amount of variability in accuracy by crime type. Specifically, 5% of homicide locations geocoded against NAVTEQ data have a positional error greater than 216 meters, compared to 268 meters for drug offense locations—a range of 52 meters. Five percent of burglary locations geocoded against Tele Atlas data have a positional error greater than 326 meters, but over 1 kilometer for homicide locations—a range of nearly 750 meter. Collectively, the findings summarized above have important implications for both the spatio-temporal analysis of crime as well as on future research.

Match rates associated with local centerlines consistently exceeded 90% and overall perform better than other reference data, including commercial data. Although previous research has demonstrated local centerline reference data can outperform other "free" street network data[17] (see, for example, Zandbergen, 2011), prior to these results no known evidence exists that suggests local centerline reference data also performs as good—if not better than—commercial reference data. These findings suggest that reference data required to produce geocoded crime incidents successfully and of high quality do not necessarily mean a large financial investment on the part of law enforcement agencies or researchers interested in the geospatial analysis of crime. Although these data are typically purchased and/or developed by a local agency (i.e., a city or county GIS department), these agencies often make the data available for free or at a minimal cost to the public, academics, and/or other local entities. However, high-quality local centerline data might not be available for all law enforcement jurisdictions, especially for those located in small, rural areas and that do not have a local GIS department that can make these data available.

Disparities in match rates are most pronounced for crime event locations where less detailed input address information are available (i.e., Intersections only), and differences in match rates among these events vary dramatically by crime type. This is likely due to the nature of the information contained in police records that is used to geocode crime locations. More detailed information—including the specific and precise address of an incident—is typically contained in the address field used to geocode burglaries, for example, than other types of crime because these events often occur in and around an addressable structure. When this detailed information is missing, geocoding quality suffers. However, the overall affect of input address quality is less

[17] Older TIGER data has known quality issues, but the latest version of the improved TIGER data was used in the current study (i.e., TIGER/Line® 2009). Although these data did not perform much worse than other data sources, findings are consistent with previous research that suggests they generally lag somewhat behind street centerlines as far as quality of the results (Zandbergen, 2011).

pronounced for crime types where an addressable structure is not likely associated with the event (i.e., auto theft or robbery). Therefore, analysts and researchers conducting geospatial analysis of crime should take steps to increase the overall quality of input address information in general and for specific types of crime that should have reliable address information associated these events.

In general, the difference between geocoding techniques and the differences between crime types are similar in magnitude when geocoding quality is measured in terms of positional accuracy. For example, the positional error for the most accurate geocoding method (local street centerlines) for the lowest accuracy crime type (drug offenses) is similar to the positional error for the lowest accuracy geocoding method (StreetMap™ USA or Tele Atlas for percentiles) for the highest accuracy crime type (assaults and burglaries, respectively). This suggests both factors are equally important in characterizing the positional error of street geocoded crime events. Despite the importance of the current study, it is not without some limitations.

4.5. Limitations

First, data used in this investigation is limited to six geographic areas. Although attempts were made to include agencies that represent a mix of urban and rural locations, specific match rates and positional error results may be unique to these study areas. Nevertheless, patterns in the results are relatively similar across study areas and therefore likely have broader applicability to other agencies.

Second, our assessment of the quality of geocoding crime events focused primarily on completeness and positional accuracy. Although one dimension of repeatability of geocoding crime data was examined (i.e., using different types of reference data), effects of changes in user-defined parameters considered during the geocoding process (i.e., spelling sensitivity, minimum candidate score, and the exclusion of tied candidates) and the influence of variations in matching algorithms of different geocoding software were not investigated. Instead, these variables were held constant across all techniques.

Finally, an address point/parcel reference location for every crime event could not be produced for the positional accuracy analysis. It is likely that the crime events that did produce an address point/parcel match represent relatively higher quality input address information. If so, then estimates of positional error may be lower bound estimates since the general quality of the records not geocoded is likely lower.

4.6. Future Research

Although the limitations described above may restrict our inferences about the effects of reference data on geocoding quality of crime event data, these problems do not limit the importance of current findings, the study's contribution to the literature, or the implications it has for future research. For example, about 9% of the data used in the current study contain input address information that is described only as an intersection or where the house number is described in terms of a 100-block, which is a relatively high figure compared datasets used to study geocoding quality in other disciplines. Future research should focus on ways to improve the overall quality of input address information for crime events, especially for those types of crimes whose geocoding quality is more adversely impacted during the geocoding process by the overall quality of the input address information (i.e., burglary). Relatedly, the current study assumes the "true" location of a crime incident is best represented by the most positionally accurate reference data (i.e., address point/parcel centroid reference data). Future research should consider investigating this assumption in order to determine whether street centerline data produces better geocoding results for certain types of crimes (i.e., non-premise crime).

In addition, user-defined parameters settings contained within the address locator services were held constant in the current study. As a result, the impact of variations in spelling sensitivity, minimum candidate score, street offset, and intersection connectors, for example, on the quality of geocoded crime data remains largely unknown. Future research should therefore examine how the manipulation of these settings influences results of geocoded crime data and subsequent analysis based on these procedures.

Finally, errors in geocoded addresses and positional accuracy may adversely affect spatial analytic methods. Nevertheless, as noted above, very few studies in general and within the crime literature in particular have tried to determine the effect of geocoding quality on the results of crime analysis. Those that have (i.e., Brimicombe et al., 2007; Harada & Shimada, 2006; Ratcliffe, 2004), fail to acknowledge patterns of ungeocoded crime data and the bias that may be introduced in analyses as a result. Therefore, given the current findings, future research should consider how variations in geocoding quality impacts various crime analysis techniques such as cluster and trend analysis as well as hot spot detection and crime prediction.

In short, although results of the current study suggest characteristics of geocoded crime data possess some similarities to data used in other disciplines, many unique differences are also demonstrated. Therefore, it is imperative that the methodological literature within the field of criminology keeps pace with the growing interest in and use of geospatial techniques associated with crime analysis.

4.7. Recommendations

Collectively, results from the current study suggest that geocoding quality is affected by variations in crime type as well as reference data used during the geocoding process. Our study concludes with a series of recommendations that practitioners and academics involved in the spatio-temporal analysis of crime events should implement in order to increase the completeness and positional accuracy of street geocoded crime events.

1. **Assess the overall quality of input address information prior to geocoding**

   If a considerable number of incident records to be geocoded are associated with intersections or 100-block addresses, consider geocoding these records separately.

Results of the current study suggest that regardless of what type of reference data used in the geocoding process, lower-quality input data will produce lower-quality geocoding results.

2. **Disaggregate crime incidents and geocode like crime events separately**

Results of the current study show that the overall quality of input address information is related to crime type. This is because some crime incidents (i.e., burglaries) are more likely to be associated with more detailed address information than other types of crime (i.e., drug offenses). Therefore, one approach to improving the quality of input address information is to disaggregate crime incident data and geocode like crime types together and separate from dissimilar events.

3. **Tailor geocoding procedures to fit specific needs**

Although match rate results were generally consistent across each jurisdiction (see Tables A1 through A6 in Appendix A), some important differences are observed. Varying results can be attributed to some factors (i.e., rural versus urban jurisdictions); but other factors are not addressed in the current study. For example, unique local road names, unusual name styles, and non-standardized road types can all produce poor geocoding results.

Results can be greatly improved by customizing the geocoding process by modifying an address locator's default classification and pattern files in such a way as to account for local idiosyncrasies. ESRI's Geocoding Development Kit (GDK) provides the tools to make these customizations in ArcGIS and have been show to improve match rates substantially (Johnson, 2007). Therefore, when possible, geocoding procedures should be tailored to specific study areas/jurisdictions.

4. **Geocode to local street centerline reference data, if it is available**

Findings from the current study suggest that local centerline reference data performs as good—if not better than—commercial reference data. Results are less persuasive for more rural jurisdictions (i.e., San Diego County); but collectively suggest that if local street centerline reference data are available, then they should be used as a low-cost way to maximize geocoding quality. Of course, this assumes that more positionally accurate reference data (i.e., address point and/or parcel reference data) are not available to use in the geocoding process. Finally, if none of the aforementioned reference data are available, commercial street centerline data generally offer a fairly reliable alternative.

5. **Characterize positional accuracy prior to additional analysis**

Results show that positional accuracy of street geocoding is influenced by the type of street reference data and by crime type. It is unclear to what degree positionally inaccurate crime data has on various crime analysis techniques (i.e., hot spot detection and crime prediction). Given that many of these approaches are based on individual/aggregate measures of relative distance, spatially accurate information is paramount. Unless geocoded crime events are determined to be positionally accurate prior to conducting more robust analysis, findings from these procedures will likely misrepresent the true nature and extent of the problem at hand. Therefore, our final recommendation is that the positional accuracy of geocoding results should be determined prior to further analysis of the crime data.

5. PREDICTIVE HOTSPOT MAPPING ANALYSIS

5.1. Research Questions

We continued our investigation by examining the effects of geocoding quality on the predictive accuracy of hotspot mapping[18]. Consideration was given hotspot technique, crime type, area, as well as parameter settings. The aim of this part of our analysis answers three important questions related to predictive hotspot mapping:

1. What factors (i.e., hotspot method, crime type, and study area) influence the accuracy of predictive hotspot mapping;

2. How sensitive are predictive hotspot mapping results to parameter settings and data quality; and

3. Is there a predictive hotspot mapping technique that is most accurate?

As with the results presented in Section 4.4, answers to these questions provide researchers and practitioners with valuable guidance and insight into one of the most popular crime analysis procedures. Answers to these questions also suggest direction for future research in this area. Data and methods used for this component of our research are described in the next section.

5.2. Data and Methodology

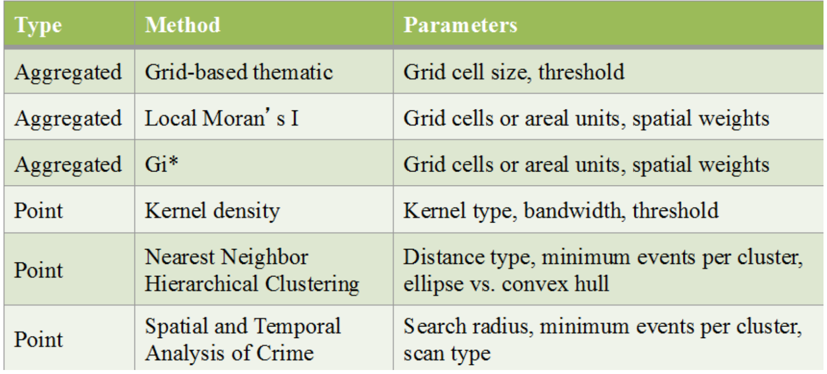
In order to answer the research questions presented in Section 5.1, we again analyzed existing crime incident data from 2007 and 2008 for the six agencies described in Section 4: Arlington Police Department (TX), Albuquerque Police Department (NM), Charlotte-Mecklenburg Police Department (NC), Las Vegas Metropolitan Police Department (NV), San Diego County Sheriff's Office (CA), and Tampa Police Department (FL).

We utilized eight different hotspot techniques, including three that are considered aggregate methods and three that are characterized as point-based methods: Grid-based thematic mapping, Local Moran's I, Gi*, Kernel Density Estimation (KDE), Nearest Neighbor Hierarchical cluster (NNH), and Spatial and Temporal Analysis of Crime (STAC). Figure 7 provides a list of the specific methods and corresponding parameter settings evaluated in the current study.

In order to answer the current research questions, all crime incident locations recorded in 2007 and 2008 for each agency were geocoded against all reference data layers (i.e., an address point reference file–or parcel centroid reference data if address point reference data were not available—and the five street reference layers described in Section 4.2). Data processing was automated using Modelbuilder and Python scripting tools in ArcGIS. Only those crime events that successfully geocoded to all reference data types were included in the current analysis. This

[18] Although the current study is focused on better understanding the effects of geocoding quality on predictive hotspot mapping, positional accuracy plays a vital role in all types of spatio-temporal analysis.

approach controlled for variations in match rates. Measures of predictive accuracy were computed, including 1) the hit rate, 2) the Predictive Accuracy Index (PAI), and 3) the Recapture Rate Index (RRI).



Type	Method	Parameters
Aggregated	Grid-based thematic	Grid cell size, threshold
Aggregated	Local Moran' s I	Grid cells or areal units, spatial weights
Aggregated	Gi*	Grid cells or areal units, spatial weights
Point	Kernel density	Kernel type, bandwidth, threshold
Point	Nearest Neighbor Hierarchical Clustering	Distance type, minimum events per cluster, ellipse vs. convex hull
Point	Spatial and Temporal Analysis of Crime	Search radius, minimum events per cluster, scan type

Figure 8. Aggregated and point pattern analysis methods and corresponding parameter settings that were used in the current study to examine the effects of geocoding quality on predictive hotspot mapping.

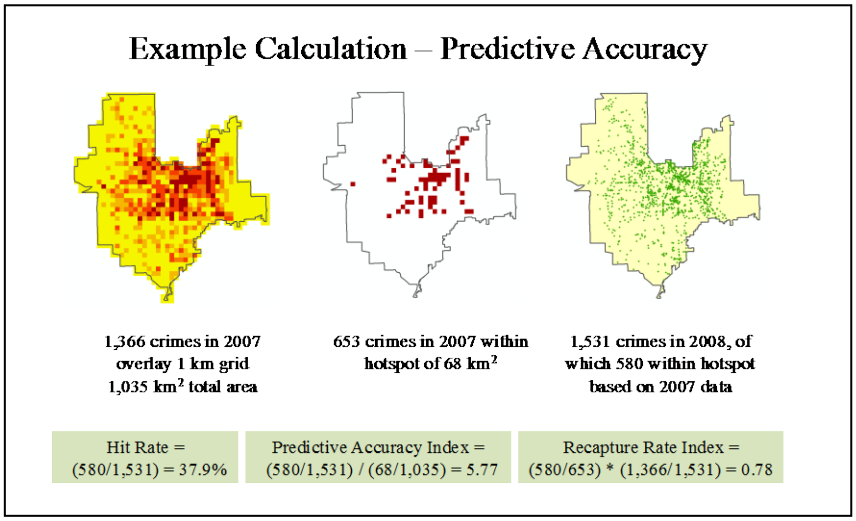
5.2.1. *Measures of predictive accuracy*. The first measure of predictive accuracy used in the current analysis was the hit rate or the percentage of crime incidents in 2008 that fell within the hotspots produced from 2007 data. A higher hit rate corresponds to greater predictive accuracy (See Figure 8).

In addition to the hit rate, the Predictive Accuracy Index (PAI) was used to measure the predictive accuracy (Chainey et al., 2008a) of hotspot maps. This provides one measure of how reliable a retrospective hotspot is able to predict future crime events, relative to the size of the hotspots. PAI is calculated as the ratio of the hit rate to the proportion of the study area that is a 2007 hotspot. Again, a higher value reflects greater accuracy.

Finally, the Recapture Rate Index (RRI) was used to determine the quality of hotspot prediction (Levine, 2008); and is based on the ratio of hotspot density for 2008 and 2007, standardized for changes in the total number of crimes in each year.

It is important to note that one important difference between the PAI and RRI is that the PAI uses the size of the hotspots (i.e., the hotspot area) in its calculation, whereas the RRI does not; and as our findings will show, is highly affected by variations in the unit of analysis size parameter setting associated with the different prediction techniques examined.

5.2.2. *Hotspot techniques*. Hotspot analysis techniques examined in the current study vary in their assumption and limitations, but what they have in common is that they rely on one or more user-defined parameters. For example, most clustering techniques require specifying the minimum number of events considered a cluster. Methods that rely on test of spatial autocorrelation require the construction of a spatial weights matrix. And KDE relies on specifying a bandwidth, among other parameters. The comparison of hotpot techniques is, therefore, sensitive to the selection of these parameters and some of the disagreement over which techniques performs best (e.g., Chainey et al., 2008 versus Levine, 2008) may in fact partly be due to differences in parameter selection.



Example Calculation – Predictive Accuracy

1,366 crimes in 2007
overlay 1 km grid
1,035 km^2 total area

653 crimes in 2007 within
hotspot of 68 km^2

1,531 crimes in 2008, of
which 580 within hotspot
based on 2007 data

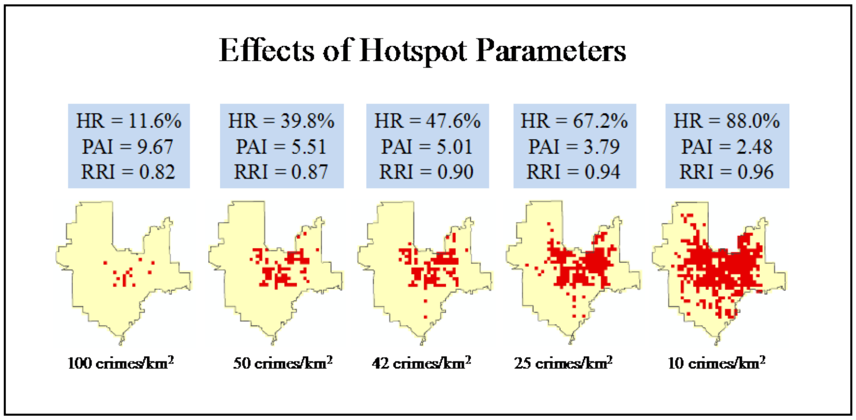
Hit Rate =
(580/1,531) = 37.9%

Predictive Accuracy Index =
(580/1,531) / (68/1,035) = 5.77

Recapture Rate Index =
(580/653) * (1,366/1,531) = 0.78

Figure 9. Examples of how the hit rate, Predictive Accuracy Index (PAI) and Recapture Rate Index (RRI) are calculated.

Analysis of the robustness of predictive hotspot mapping techniques, therefore, needs to consider the variability in these parameters. For each of the techniques presented above, a range of parameters values was employed and the validation of hotspots was repeated (see Figure 7). Results presented in Section 5.4 provide a set of parameters to achieve the best performing hotspot. However, the analytic objective was not necessarily to determine the "optimum" parameters, but to determine how the accuracy of hotspot prediction depends on critical parameters (see Research Question #2). Figure 9 illustrates the effects of parameter settings on measures of predictive accuracy used in a point-pattern analysis, where the number of events used to define a "cluster" varies.



Effects of Hotspot Parameters

HR = 11.6%
PAI = 9.67
RRI = 0.82
	HR = 39.8%
	PAI = 5.51
	RRI = 0.87
		HR = 47.6%
		PAI = 5.01
		RRI = 0.90
			HR = 67.2%
			PAI = 3.79
			RRI = 0.94
				HR = 88.0%
				PAI = 2.48
				RRI = 0.96

100 crimes/km^2	50 crimes/km^2	42 crimes/km^2	25 crimes/km^2	10 crimes/km^2

Figure 10. An illustration of how variation in the parameter setting that defines a "cluster" affects predictive accuracy (i.e., hit rate (HR), the Predictive Accuracy Index (PAI), and the Recapture Rate Index (RRI)).

5.3. Results

In order to answer the research questions presented in Section 5.1, the predictive accuracy of crime hotspot mapping was analyzed by technique, crime type, and study area. Variations in select parameter settings were also considered in the context of geocoding quality. Results of these analyses are presented below, respectively; and provide insight into which predictive mapping techniques are most accurate.

5.3.1. *Hotspot method*. As noted above, a variety of methods can be used to produce predictive crime hotspot maps. Therefore, our predictive hotspot mapping analysis began by considering the predictive accuracy of these techniques, while holding relative parameter settings constant across methods. This analysis was also based on a single geocoding technique—street geocoded data—since variations in the quality of geocoded data might also affect results. Results are presented in Table 4. Findings show that the influence of hotspot technique on predictive accuracy, measured in terms of hit rate, PAI, and RRI, varies substantially across techniques.

First, the hit rate—a proportion of 2008 crime incident locations that falls within 2007 hotspots—varies from a high of over 47% for Local Moran's I (aggregating crime incidents to the block group) to a low of just over 7% for the grid-based thematic mapping technique. Second, considerable variation in the PAI and RRI was also observed across mapping techniques. PAI was relatively more volatile than the RRI. This was not surprising, however, since PAI considers the total area associated with the 2007 hotspots, relative to the overall size of the study area; but RRI does not. Collectively, these findings suggest that there are trade-offs among accuracy metrics, but regardless of metric used the predictive accuracy of crime hotspot maps are strongly influenced by the approach used to generate them.

Since these results were based on data aggregated, both in terms of crime type and by urban morphology, which could explain some of the variation across predictive metrics, we considered the effects of each of these factors, independently; and results of these analyses are presented next.

5.3.2. *Crime type.* Table 5 presents results of our analysis of predictive accuracy when crime type and hotspot method were considered independently. As with the previous analysis, both parameter settings and geocoding quality were held constant. Findings show that crime type has some effect on predictive accuracy of hotspot mapping techniques, but that the effect is not as substantial as technique alone. For example, among most techniques, predictive accuracy is considerably lower for homicide than for other crime types. This finding is likely due to the relatively fewer number of homicides than other types of crime. Another noteworthy finding is that the PAI is highest for the two hierarchical techniques (i.e., NNH and STAC). A similar pattern was not observed for the RRI, which suggests that not only does crime type play an important role in the predictive accuracy of crime hotspot mapping, but that the consideration of study area is more pronounced when data is disaggregated by crime type. Given these findings, urban morphology was a factor also considered to affect predictive accuracy.

Table 4. The influence of hotspot method on predictive accuracy for street geocoded crimes within all jurisdictions.

	Crimes in 2007–		Crimes in 2008–		Total (in 1,000m ^{2})–		Predictive accuracy		
Hotspot method	in 2007 hotspots	in study area	in 2007 hotspots	in study area	area of 2007 hotspots	study area	Hit rate (%)	PAI	RRI
Grid-based Thematic Mapping[1]	18,135	174,889	12,422	174,300	106,903	86,155,518	7.13	57.44	0.69
Local Moran's I - Grid[2]	96,942	166,413	67,451	162,347	1,004,438	86,155,518	41.55	35.64	0.71
Local Moran's I - Enumeration Area[3]	104,685	166,200	76,857	161,993	1,742,594	86,058,250	47.44	23.43	0.75
Gi* - Grid[4]	52,181	165,184	44,438	160,841	1,330,610	85,571,909	27.63	17.77	0.87
Gi* - Enumeration Area[5]	84,549	166,200	70,070	161,993	2,657,449	86,058,250	43.25	14.01	0.85
Kernel Density Estimation[6]	41,371	165,184	33,390	160,841	928,252	85,571,909	20.76	19.14	0.83
NNH[7]	76,218	166,413	47,667	162,347	404,306	86,155,518	29.36	62.57	0.64
STAC[8]	26,562	166,413	18,982	162,347	22,094	86,155,518	11.69	455.93	0.73

[1] Based on 250m grid cell size and threshold set at 20%.

[2] Based on 250m grid cell size, IDW squared threshold set at 0, with no row standardization.

[3] Aggregated to the block group, IDW squared threshold set at 0, with no row standardization.

[4] Based on 250m grid cell size, fixed distance set at 37m, with row standardization.

[5] Aggregated to the block group, fixed distance set at 37m, with row standardization.

[6] Search radius set at 200m, threshold set at more than 3 times the mean.

[7] Search radius set at 200m, using covex hull visualization with a minimum cluster threshold of of 15 points. First-order clustering was utilized.

[8] Search radius set at 200m, using covex hull visualization with a minimum cluster threshold of of 15 points.

Table 5. The influence of crime type on predictive accuracy for street geocoded incidents, by hotspot method.

	Predictive accuracy		
Hotspot method and crime type	Hit rate (%)	PAI	RRI
Grid-based Thematic Mapping[1]			
Assault	33.38	40.55	0.64
Auto Burglary	42.72	27.73	0.74
Auto Theft	42.88	34.68	0.76
Burglary	34.43	19.97	0.68
Drug Offenses	45.74	63.84	0.77
Homicide	1.63	71.16	0.07
Robbery	35.68	72.30	0.72
Local Moran's I - Grid[2]			
Assault	51.44	21.19	0.61
Auto Burglary	45.55	17.32	0.75
Auto Theft	52.25	18.22	0.77
Burglary	47.62	12.86	0.74
Drug Offenses	59.40	32.47	0.80
Homicide	3.66	57.18	0.07
Robbery	61.20	34.80	0.74
Local Moran's I - Enumeration Area[3]			
Assault	29.60	22.46	0.86
Auto Burglary	20.91	16.73	0.87
Auto Theft	24.62	18.12	0.86
Burglary	20.80	13.02	0.86
Drug Offenses	34.74	28.17	0.90
Homicide	8.50	21.25	0.24
Robbery	31.78	24.19	0.89
Gi* - Grid[4]			
Assault	50.84	13.68	0.85
Auto Burglary	40.71	10.38	0.84
Auto Theft	47.67	12.66	0.87
Burglary	43.92	7.89	0.85
Drug Offenses	57.70	21.13	0.93
Homicide	16.67	17.18	0.17
Robbery	57.83	17.51	0.87
Gi* - Enumeration Area[5]			
Assault	23.62	17.56	0.91
Auto Burglary	17.79	14.68	0.95
Auto Theft	18.77	15.55	0.89
Burglary	16.09	12.17	0.91
Drug Offenses	31.78	24.87	0.90
Homicide	13.77	15.87	0.40
Robbery	25.84	21.62	0.97

(continued)

Table 5. The influence of crime type on predictive accuracy for street geocoded incidents, by hotspot method (continued).

	Predictive accuracy		
Hotspot method and crime type	Hit rate (%)	PAI	RRI
Kernel Density Estimation[6]			
Assault	27.14	52.65	0.53
Auto Burglary	35.68	41.93	0.70
Auto Theft	35.76	52.42	0.71
Burglary	26.15	26.32	0.56
Drug Offenses	40.03	96.26	0.72
Homicide	0.41	168.23	0.05
Robbery	34.06	106.02	0.64
NNH[7]			
Assault	4.67	252.82	0.43
Auto Burglary	15.69	419.66	0.70
Auto Theft	15.97	377.01	1.08
Burglary	8.67	169.57	0.66
Drug Offenses	18.44	519.44	0.72
Homicide	-	-	-
Robbery	4.76	808.40	0.69
STAC[8]			
Assault	2.13	293.71	0.48
Auto Burglary	4.01	411.14	0.68
Auto Theft	4.09	853.98	0.68
Burglary	1.80	172.63	0.60
Drug Offenses	6.51	493.22	0.70
Homicide	-	-	-
Robbery	1.67	761.29	0.62

[1] Based on 250m grid cell size and threshold set at 20%.

[2] Based on 250m grid cell size, IDW squared threshold set at 0, with no row standardization.

[3] Aggregated to the block group, IDW squared threshold set at 0, with no row standardization.

[4] Based on 250m grid cell size, fixed distance set at 37m, with row standardization.

[5] Aggregated to the block group, fixed distance set at 37m, with row standardization.

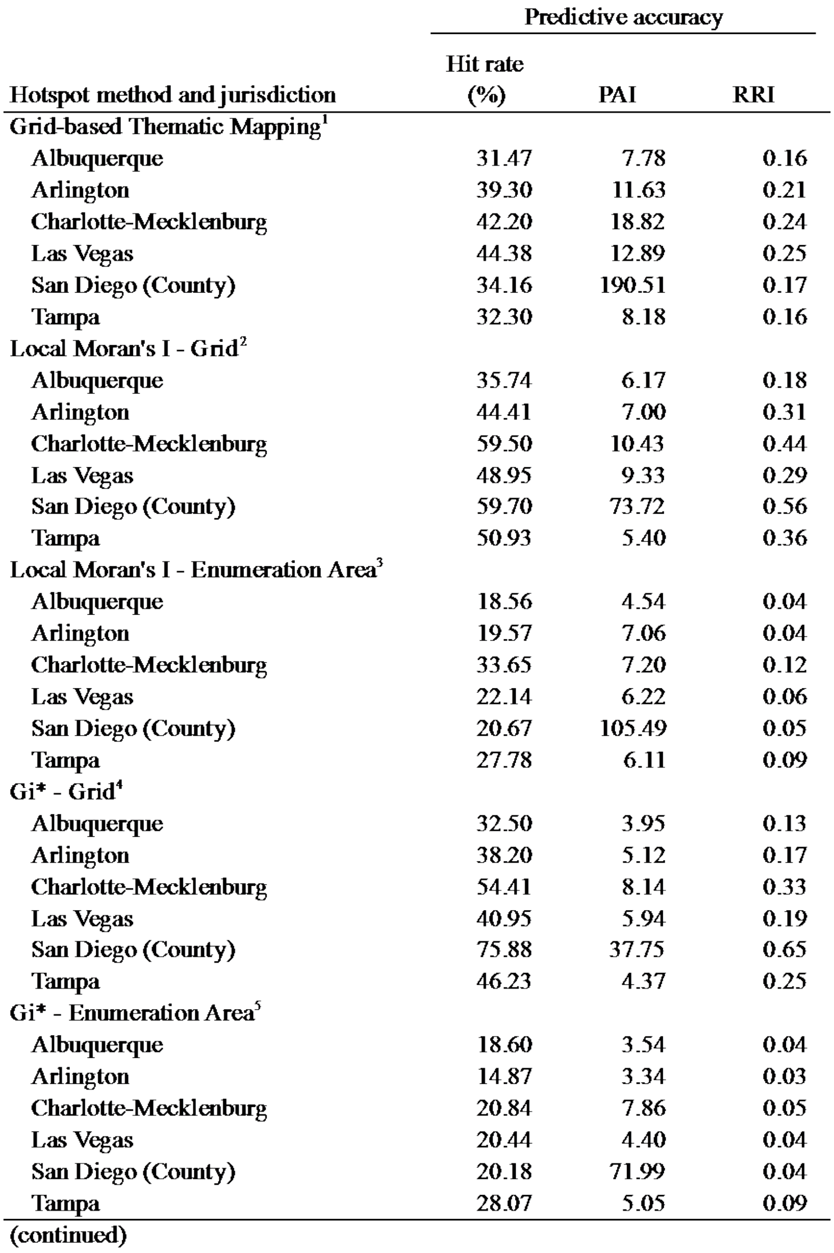
[6] Search radius set at 200m, threshold set at more than 3 times the mean.

[7] Search radius set at 200m, using covex hull visualization with a minimum cluster threshold of of 15 points. First-order clustering was utilized.

[8] Search radius set at 200m, using covex hull visualization with a minimum cluster threshold of of 15 points.

5.3.3. *Study area.* In addition to mapping technique and crime type, we examined the effects of variation in study area on the predictive accuracy of hotspot mapping techniques. Results are presented in Table 6 and show that the influence of urban morphology on predictive accuracy has a modest effect for each hotspot method considered. For example, Arlington, Albuquerque, and Tampa are jurisdictions consistently associated with the lowest predictive accuracy scores, regardless of metric. And although there are some differences observed in the hit rates and the RRI between study areas, more modest differences in the PAI are apparent. Specifically, San

Table 6. The influence of urban morphology on predictive accuracy for street geocoded incidents, by hotspot method.



	Predictive accuracy		
Hotspot method and jurisdiction	Hit rate (%)	PAI	RRI
Grid-based Thematic Mapping[1]			
Albuquerque	31.47	7.78	0.16
Arlington	39.30	11.63	0.21
Charlotte-Mecklenburg	42.20	18.82	0.24
Las Vegas	44.38	12.89	0.25
San Diego (County)	34.16	190.51	0.17
Tampa	32.30	8.18	0.16
Local Moran's I - Grid[2]			
Albuquerque	35.74	6.17	0.18
Arlington	44.41	7.00	0.31
Charlotte-Mecklenburg	59.50	10.43	0.44
Las Vegas	48.95	9.33	0.29
San Diego (County)	59.70	73.72	0.56
Tampa	50.93	5.40	0.36
Local Moran's I - Enumeration Area[3]			
Albuquerque	18.56	4.54	0.04
Arlington	19.57	7.06	0.04
Charlotte-Mecklenburg	33.65	7.20	0.12
Las Vegas	22.14	6.22	0.06
San Diego (County)	20.67	105.49	0.05
Tampa	27.78	6.11	0.09
Gi* - Grid[4]			
Albuquerque	32.50	3.95	0.13
Arlington	38.20	5.12	0.17
Charlotte-Mecklenburg	54.41	8.14	0.33
Las Vegas	40.95	5.94	0.19
San Diego (County)	75.88	37.75	0.65
Tampa	46.23	4.37	0.25
Gi* - Enumeration Area[5]			
Albuquerque	18.60	3.54	0.04
Arlington	14.87	3.34	0.03
Charlotte-Mecklenburg	20.84	7.86	0.05
Las Vegas	20.44	4.40	0.04
San Diego (County)	20.18	71.99	0.04
Tampa	28.07	5.05	0.09

(continued)

Table 6. The influence of urban morphology on predictive accuracy for street geocoded incidents, by hotspot method (continued).

	Predictive accuracy		
Hotspot method and jurisdiction	Hit rate (%)	PAI	RRI
Kernel Density Estimation[6]			
Albuquerque	26.22	14.50	0.12
Arlington	32.16	20.68	0.15
Charlotte-Mecklenburg	33.01	17.03	0.19
Las Vegas	37.91	24.00	0.19
San Diego (County)	30.78	328.96	0.15
Tampa	24.81	13.99	0.11
NNH[7]			
Albuquerque	7.83	80.66	0.01
Arlington	18.94	111.13	0.05
Charlotte-Mecklenburg	6.98	1,065.68	0.01
Las Vegas	20.49	124.54	0.05
San Diego (County)	7.01	2,157.58	0.01
Tampa	8.19	39.18	0.01
STAC[8]			
Albuquerque	3.05	80.45	0.00
Arlington	3.81	112.21	0.00
Charlotte-Mecklenburg	1.65	2,212.22	0.00
Las Vegas	5.21	208.25	0.00
San Diego (County)	2.65	2,613.65	0.00
Tampa	3.18	34.72	0.00

[1] Based on 250m grid cell size and threshold set at 20%.

[2] Based on 250m grid cell size, IDW squared threshold set at 0, with no row standardization.

[3] Aggregated to the block group, IDW squared threshold set at 0, with no row standardization.

[4] Based on 250m grid cell size, fixed distance set at 37m, with row standardization.

[5] Aggregated to the block group, fixed distance set at 37m, with row standardization.

[6] Search radius set at 200m, threshold set at more than 3 times the mean.

[7] Search radius set at 200m, using covex hull visualization with a minimum cluster threshold of of 15 points. First-order clustering was utilized.

[8] Search radius set at 200m, using covex hull visualization with a minimum cluster threshold of of 15 points.

Diego County is associated with the highest predictive accuracy scores as measured by the PAI; however, this is likely to the use of the total size of a study area in the computation of the PAI because San Diego County is the largest and least densely populated study area examined.

To summarize, thus far, our analysis shows that hotspot technique substantially affects predictive accuracy, crime type has a moderate effect on predictive accuracy, and study area has a modest effect on predictive accuracy. For each technique considered in our analysis, a number of parameters must be defined prior to the production of any hotspot map. Variations in these settings are also believed to affect the predictive accuracy of results; therefore, we continue our analysis by examining the effects of user-defined parameter settings on predictive accuracy in order to answer the second research question listed in Section 5.1.

5.3.4. *Parameter settings.* One of the most popular predictive hotspot mapping techniques used in crime analysis is Kernel Density Estimation (KDE). The popularity of KDE for the production of predictive crime hotspot maps is due in part to the belief that it is “the most suitable spatial analysis technique for visualizing crime data” (Chainey et al., 2008, p. 8). The growing availability of KDE in popular GIS applications, the perceived accuracy of its hotspot identification, and the aesthetically pleasing and easily understandable output are other factors that have lead to its popularity.

The mechanics of any point pattern analysis used to detect crime hotspots (i.e., KDE, NNH, STAC) are similar in many ways. They each calculate the density of crime events in a given area based on incident locations in surrounding areas and produce a raster output based on a grid overlay. These calculations rely on mathematical formulas that are influenced by several factors, including the form of the curved surface that is fit over each point, the cell size of the grid overlay, and the length of the search radius used to identify neighboring crime events.

Given the widespread adoption of KDE, the effects of parameter settings on predictive accuracy described below are based solely on this technique. Furthermore, while the geocoding method and study areas are also held constant (street centerline geocoding and Charlotte-Mecklenburg, respectively), the effects of two important parameter settings (i.e., search radius and threshold size) vary across crime type so that a more robust understanding of the effects of parameter settings on predictive accuracy can be discerned.

For the results that follow, a total of ten different search radii were considered. The kernel bandwidths ranged from 50m to 500m, set at 50m intervals. Similarly, five different threshold settings were considered. Cell sizes were set at less than the average, two times the average, three times the average, four times the average, and five times the average. Figures 11-17 present results of our analysis of variations in these parameter settings when considered simultaneously for each of the eight crime types considered, respectively.

Findings of our analysis suggest that using the average cell size and a relatively low bandwidth produce an acceptable hit rate, regardless of crime type. And while there is a corresponding increase in hit rate with an increase in kernel bandwidth, the increase is not proportionate to the increase in the search radius. For example, doubling the bandwidth from 50m to 100m or from 100m to 200m does not result in a corresponding two-fold increase in hit rate.

Although our findings show that relatively lower thresholds correspond to relatively higher hit rates across crime types, using a low threshold also appears to correspond to the lowest predictive accuracy when measured in terms of PAI and RRI. For example, when KDE parameters are set at the average grid cell size and a kernel bandwidth of 50m the PAI for assaults is 61.0; but when it is increased to five times the average cell size, the PAI increases to 358.3 (see Figure 11). Similarly, when KDE parameters are set at the average grid cell size and a kernel bandwidth of 50m the PAI for burglary is 28.9.0; but when it is increased to five times the average cell size, the

PAI increases to 152.8 (see Figure 14). When the search radius is increased to between 150m to 200m, however, the influence of threshold size becomes less pronounced for most crimes.

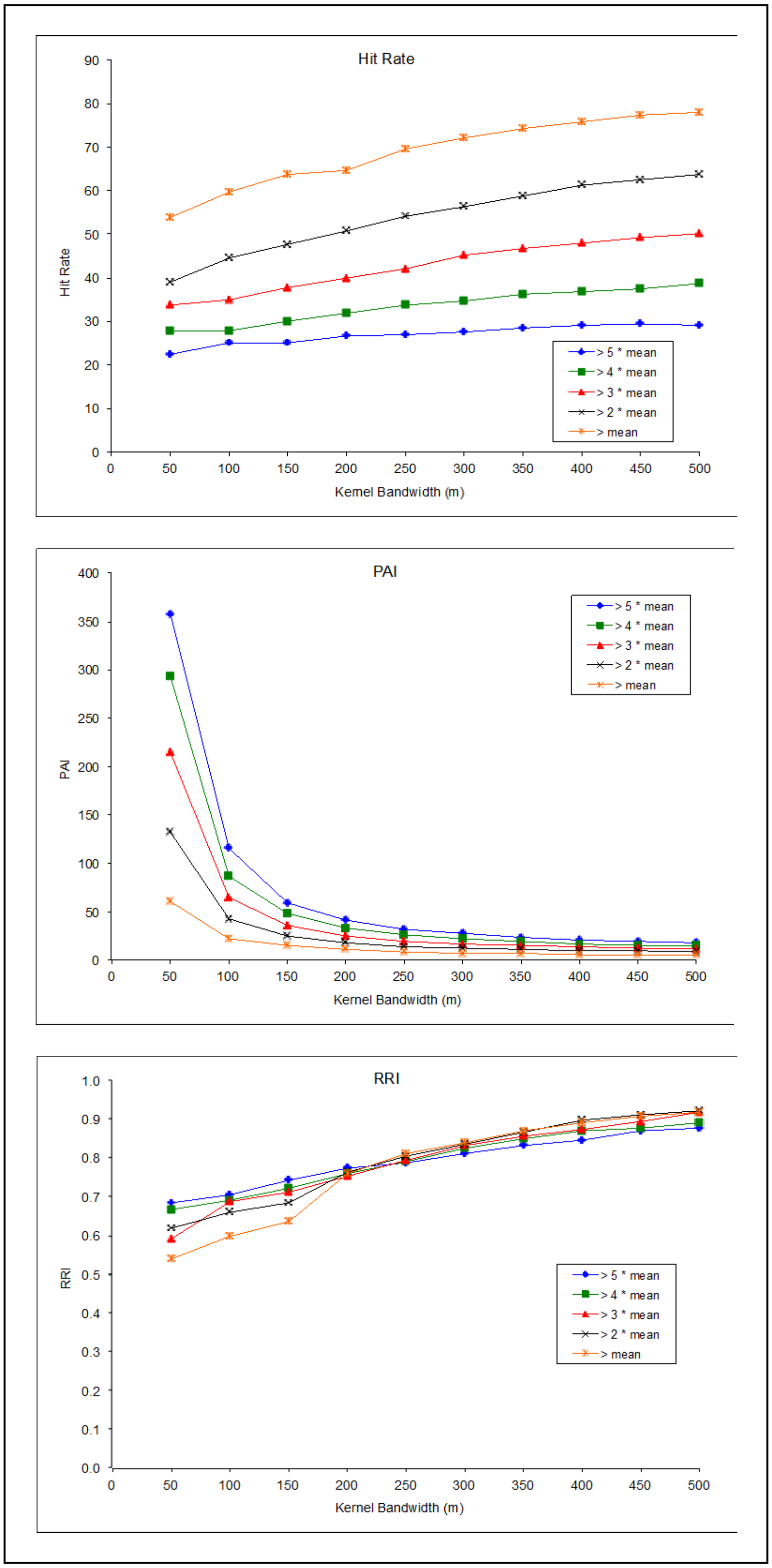


Figure 11. Predictive accuracy metrics produced from KDE. Analysis based on assaults in Charlotte-Mecklenburg, using different threshold sizes and search radii.

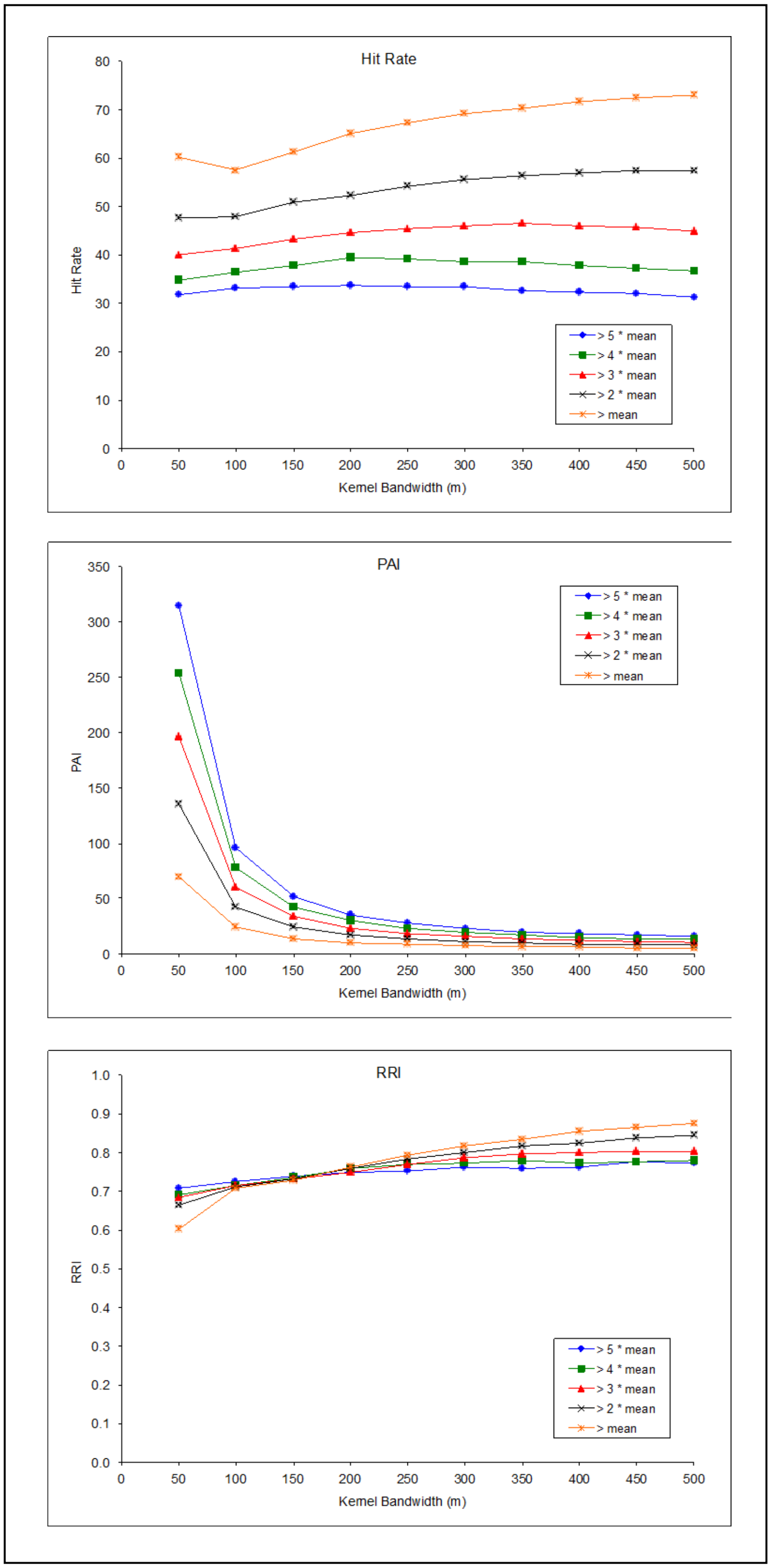


Figure 12. Predictive accuracy metrics produced from KDE. Analysis based on auto burglaries in Charlotte-Mecklenburg, using different threshold sizes and search radii.

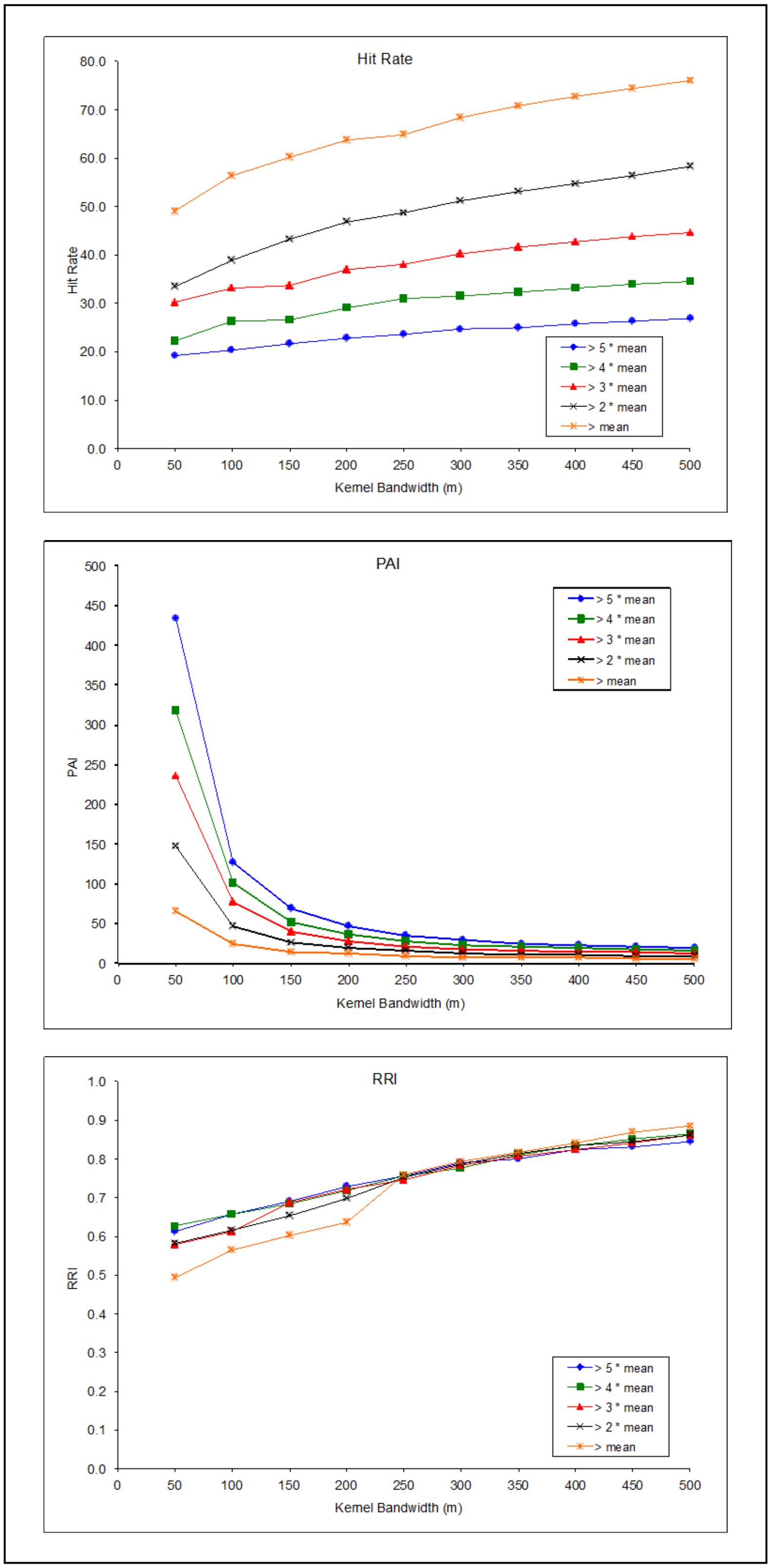


Figure 13. Predictive accuracy metrics produced from KDE. Analysis based on auto thefts in Charlotte-Mecklenburg, using different threshold sizes and search radii.

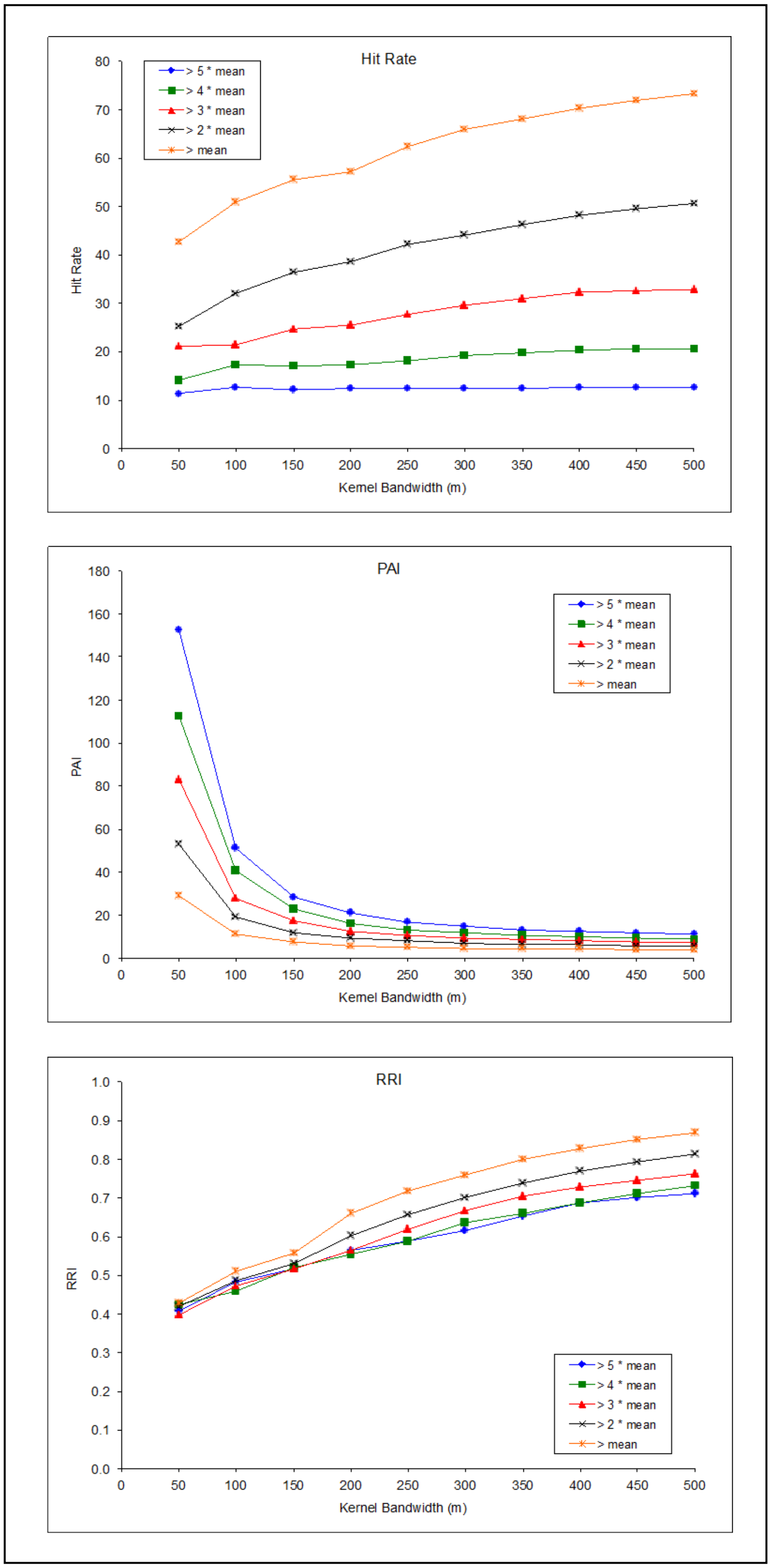


Figure 14. Predictive accuracy metrics produced from KDE. Analysis based on burglaries in Charlotte-Mecklenburg, using different threshold sizes and search radii.

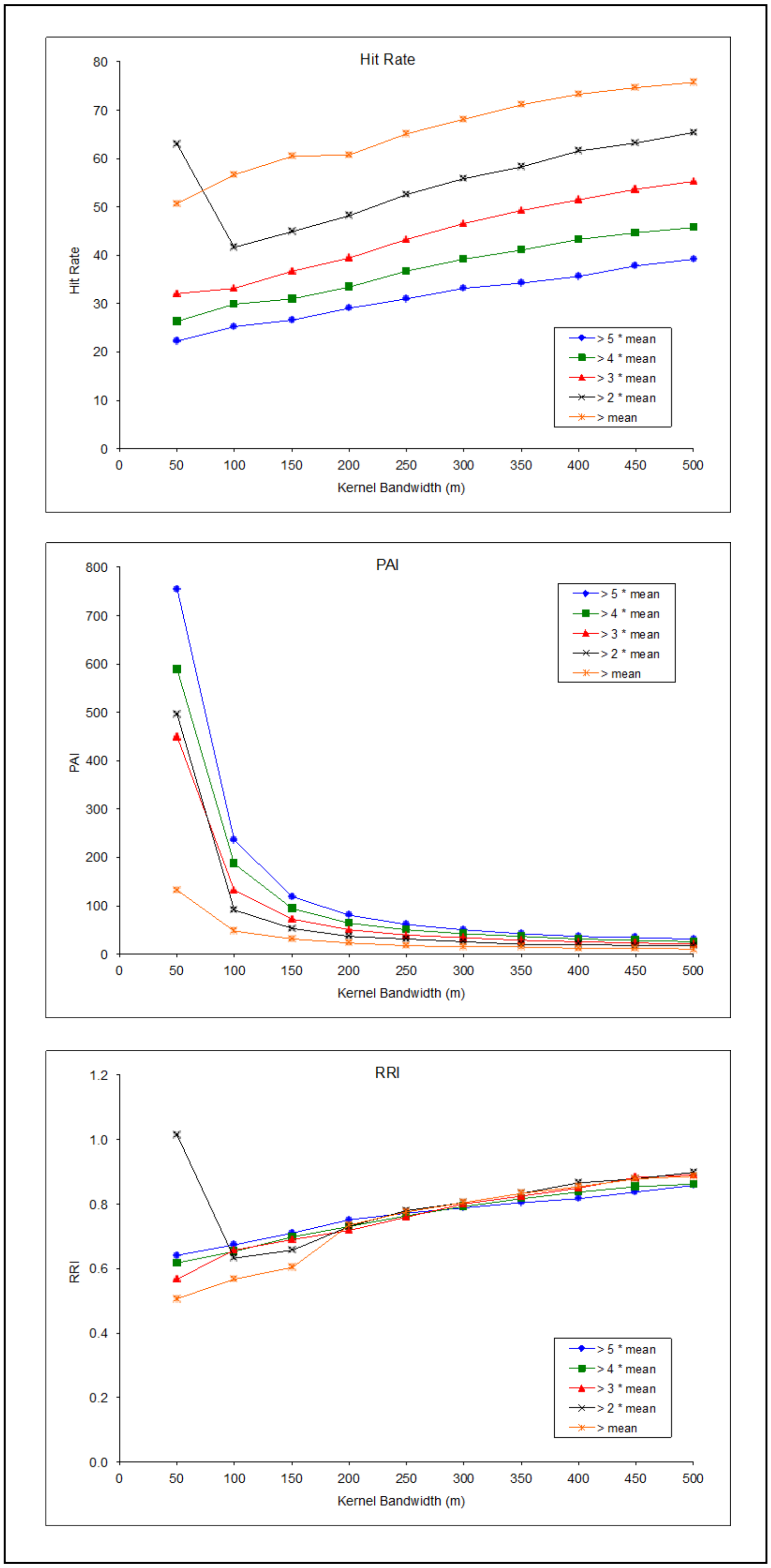


Figure 15. Predictive accuracy metrics produced from KDE. Analysis based on drug offenses in Charlotte-Mecklenburg, using different threshold sizes and search radii.

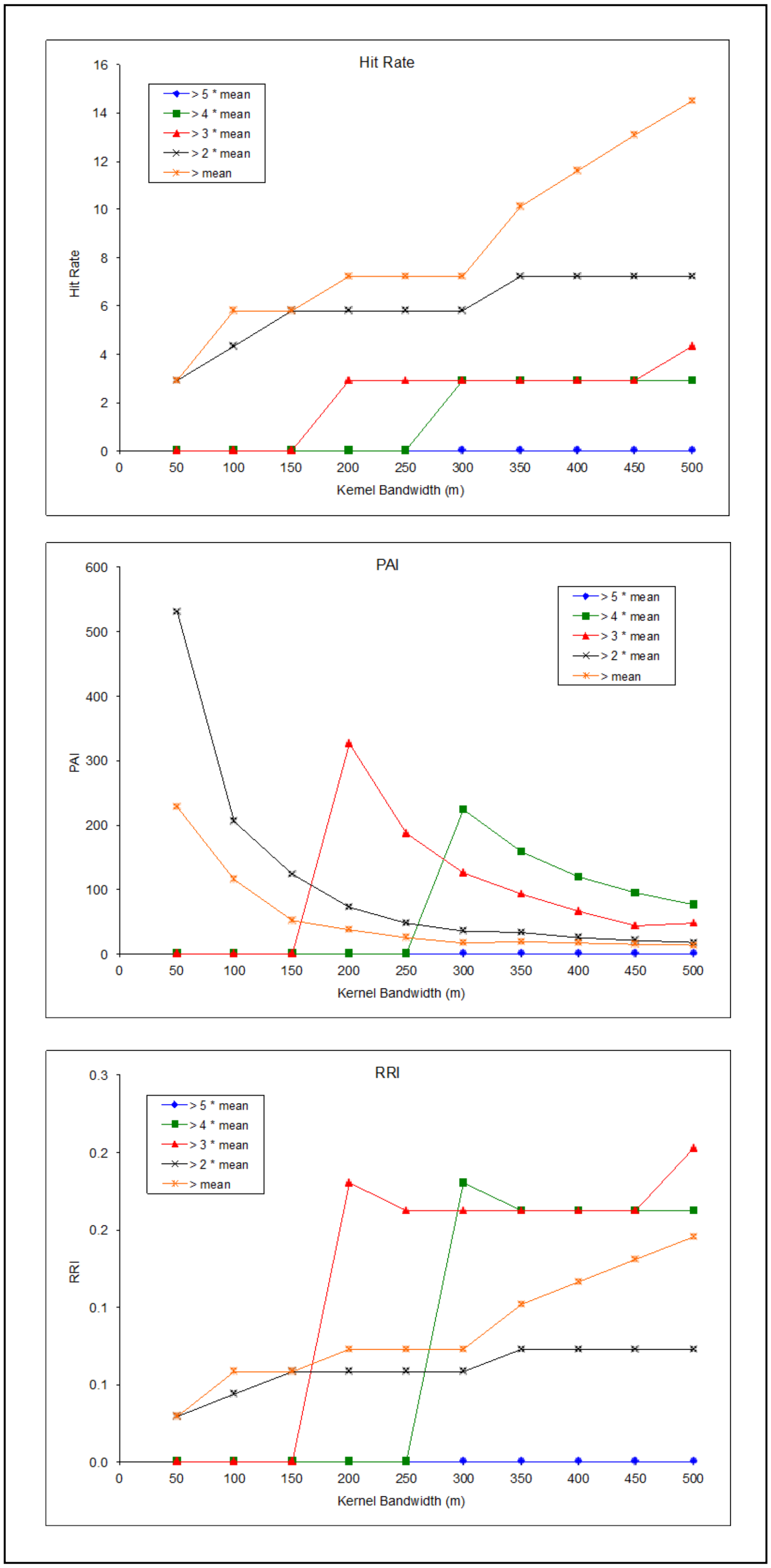


Figure 16. Predictive accuracy metrics produced from KDE. Analysis based on homicides in Charlotte-Mecklenburg, using different threshold sizes and search radii.

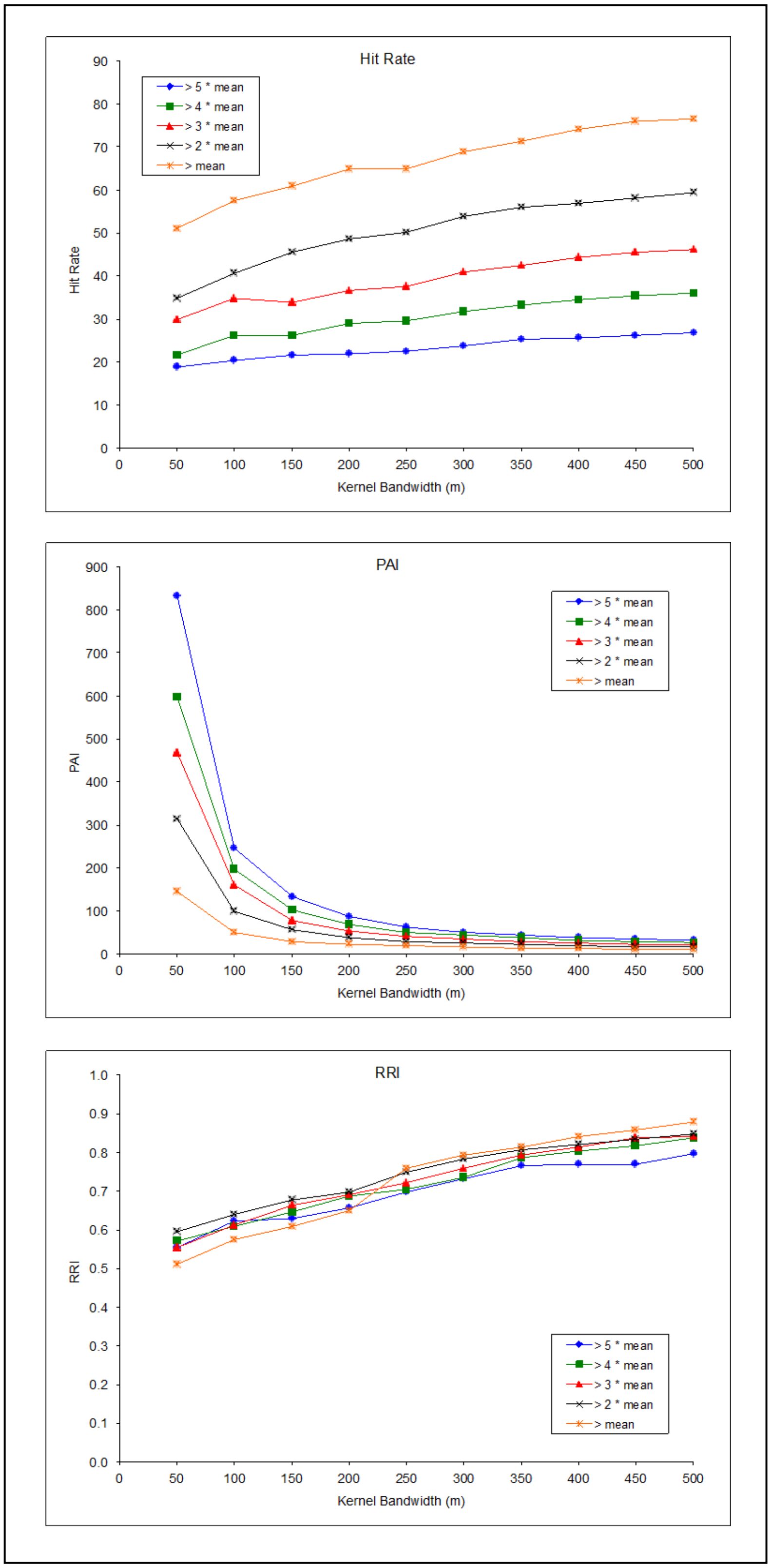


Figure 17. Predictive accuracy metrics produced from KDE. Analysis based on robberies in Charlotte-Mecklenburg, using different threshold sizes and search radii.

Finally, results show that the effects of parameter settings on predictive accuracy impact the RRI least of all. Specifically, the RRI graphs presented in Figures 11-17 not only show lines for each threshold setting that are tightly bunched together, but also show lines that rise gradually. This means that neither the effects of kernel bandwidth nor cell size have a pronounced effect on the RRI. Combined, these results suggest that parameter settings have a considerable effect on crime predictions.

5.3.5. *Geocoding quality*. Our final analysis examined the influence of geocoding quality on predictive hotspot mapping. Given the influence of technique, crime type, study area, and parameter setting on predictive accuracy that were already demonstrated, we analyzed the influence of geocoding quality while considering the influence of these other factors, simultaneously.

We began by investigating the influence of geocoding quality on predictive crime hotspot produced by a thematic grid-based technique. Maps were generated for two crime types (i.e., assaults and burglary), three study areas (i.e., Arlington, Charlotte, and San Diego (County)), and two difference cell sizes (i.e., 100m and 250m). Predictive accuracy was measured in terms of hit rate, PAI, and RRI. Results are presented in Table 7 and indicate that in general, the overall quality of geocoded data dramatically affects predictive accuracy. As cell size increases, the corresponding hit rate increases; however, a comparable proportionate increase in the hotspot area is not observed and therefore predictive accuracy as reflected in the PAI is decreased.

Table 7. The influence of cell size[1] and geocoding method on the pedictive accuracy of grid-based thematic maps of assaults and burglaries, by jurisdiction.

	Assault			Burglary		
	Predictive accuracy			Predictive accuracy		
Jurisdiction, cell size, and geocoding method	Hit rate (%)	PAI	RRI	Hit rate (%)	PAI	RRI
Arlington						
100m						
Address Points	1.46	9.42	0.05	3.48	3.23	0.09
Street Centerlines	11.22	72.21	0.38	9.64	8.95	0.25
NAVTEQ	9.76	62.79	0.33	7.47	6.94	0.19
TeleAtlas	8.29	53.37	0.28	6.44	5.99	0.17
StreetMap	7.80	50.23	0.26	7.13	6.62	0.18
TIGER	8.29	53.37	0.28	7.13	6.62	0.18
250m						
Address Points	8.29	9.23	0.24	15.68	3.17	0.35
Street Centerlines	13.66	15.21	0.39	18.87	3.81	0.42
NAVTEQ	11.71	13.03	0.34	17.62	3.56	0.39
TeleAtlas	10.73	11.95	0.31	17.62	3.56	0.39
StreetMap	10.73	11.95	0.31	18.42	3.72	0.41
TIGER	10.24	11.40	0.29	18.59	3.75	0.42

(continued)

Table 7. The influence of cell size[1] and geocoding method on the pedictive accuracy of grid-based thematic maps of assaults and burglaries, by jurisdiction (continued).

	Assault			Burglary		
	Predictive accuracy			Predictive accuracy		
Jurisdiction, cell size, and geocoding method	Hit rate (%)	PAI	RRI	Hit rate (%)	PAI	RRI
Charlotte						
100m						
Address Points	25.96	55.35	0.58	13.72	19.79	0.35
Street Centerlines	13.92	29.67	0.31	8.67	12.51	0.22
NAVTEQ	14.04	29.93	0.31	9.14	13.18	0.23
TeleAtlas	13.39	28.54	0.30	9.17	13.22	0.23
StreetMap	13.86	29.54	0.31	8.64	12.46	0.22
TIGER	13.77	29.37	0.31	8.54	12.31	0.22
250m						
Address Points	40.59	21.20	0.77	29.04	10.26	0.61
Street Centerlines	35.08	18.32	0.66	27.99	9.89	0.59
NAVTEQ	35.85	18.72	0.68	28.40	10.03	0.60
TeleAtlas	35.85	18.72	0.68	28.79	10.17	0.60
StreetMap	34.69	18.12	0.66	27.63	9.76	0.58
TIGER	34.97	18.27	0.66	28.47	10.06	0.60
San Diego (County)						
100m						
Address Points	13.07	757.02	0.39	14.26	406.67	0.38
Street Centerlines	4.27	247.34	0.13	4.52	128.98	0.12
NAVTEQ	3.75	217.36	0.11	3.57	101.67	0.09
TeleAtlas	3.10	179.89	0.09	3.78	107.74	0.10
StreetMap	4.92	284.82	0.15	3.73	106.22	0.10
TIGER	4.27	247.34	0.13	2.98	84.98	0.08
250m						
Address Points	20.05	216.40	0.52	21.71	118.49	0.50
Street Centerlines	19.53	210.82	0.50	16.50	90.03	0.38
NAVTEQ	18.24	196.85	0.47	17.30	94.39	0.40
TeleAtlas	18.76	202.44	0.48	15.97	87.13	0.37
StreetMap	15.39	166.14	0.40	16.34	89.16	0.38
TIGER	15.78	170.33	0.41	16.50	90.03	0.38

[1] Based on threshold set at 20%.

Conversely, the RRI increases as grid cell size increases. Regardless of study area, crime type, or parameter setting, the key finding of these analysis is that the use of street geocoded data does not have a consistent, negative effect on predictive accuracy.

Next, we determined the influence of grid cell size versus enumeration area (i.e., aggregating data to the block group), and geocoding method for Local Moran's I on predictive accuracy. As with the grid-based thematic mapping technique, data were analyzed for two crime types and three study areas; and are presented in Table 8. Results suggest that the influence of geocoding quality is less influential than the size of the unit of analysis (i.e., cell size parameter), study area,

Table 8. The influence of cell size[1] and geocoding method on the pedictive accuracy of Local Moran's I maps of assaults and burglaries, by jurisdiction.

Jurisdiction, cell size, and geocoding method	Assault			Burglary		
	Predictive accuracy			Predictive accuracy		
	Hit rate (%)	PAI	RRI	Hit rate (%)	PAI	RRI
Arlington						
250m						
Address Points	18.05	6.46	0.27	31.64	3.11	0.54
Street Centerlines	23.90	8.56	0.35	35.40	3.48	0.60
NAVTEQ	22.44	8.04	0.33	33.64	3.31	0.57
TeleAtlas	22.44	8.04	0.33	34.04	3.35	0.58
StreetMap	20.49	7.34	0.30	34.21	3.36	0.58
TIGER	21.46	7.69	0.32	34.38	3.38	0.59
Block groups						
Address Points	60.49	1.98	0.72	82.84	1.45	0.92
Street Centerlines	59.51	1.95	0.71	82.44	1.44	0.92
NAVTEQ	60.98	2.00	0.72	82.78	1.45	0.92
TeleAtlas	60.00	1.97	0.71	82.61	1.44	0.92
StreetMap	58.54	1.92	0.69	82.90	1.45	0.93
TIGER	59.02	1.93	0.70	81.87	1.43	0.91
Charlotte						
250m						
Address Points	67.67	10.78	0.83	60.88	6.34	0.75
Street Centerlines	63.52	10.12	0.78	58.72	6.12	0.72
NAVTEQ	63.21	10.07	0.78	59.61	6.21	0.73
TeleAtlas	63.60	10.14	0.78	59.76	6.23	0.73
StreetMap	61.53	9.81	0.76	59.25	6.17	0.73
TIGER	62.95	10.03	0.77	59.59	6.21	0.73
Block groups						
Address Points	44.15	9.71	1.00	39.99	6.00	0.88
Street Centerlines	42.99	9.45	0.97	39.37	5.91	0.87
NAVTEQ	44.23	9.72	1.00	39.97	6.00	0.88
TeleAtlas	44.11	9.70	1.00	40.08	6.01	0.88
StreetMap	44.25	9.73	1.00	39.45	5.92	0.87
TIGER	44.23	9.72	1.00	39.96	5.99	0.88
San Diego (County)						
250m						
Address Points	--	–	–	–	–	--
Street Centerlines	--	–	–	–	–	--
NAVTEQ	--	–	–	–	–	--
TeleAtlas	--	–	–	–	–	--
StreetMap	--	–	–	–	–	--
TIGER	--	–	–	–	–	--
Block groups						
Address Points	28.20	116.44	0.89	30.76	82.09	1.09
Street Centerlines	27.68	114.31	0.87	30.71	81.94	1.09
NAVTEQ	27.17	112.17	0.86	31.13	83.08	1.11
TeleAtlas	25.74	106.29	0.81	30.71	81.94	1.09
StreetMap	24.84	102.56	0.78	28.69	76.55	1.02
TIGER	27.43	113.24	0.86	30.65	81.80	1.09

[1] Based on IDW squared threshold set at 0, with no row standardization.

or crime type. That is to say, data quality appears to have little to no effect on these results, given the similarity between the predictive accuracy metrics associated with address point geocoded data and corresponding street geocoded results.

The third set of results examining the influence of geocoding quality on predictive hotspot mapping are based on the Getis-Ord Gi* technique. As with the previous analysis, cell size varies between 250m and block group enumeration areas. Again, data were analyzed for two crime types and three study areas.

Findings are presented in Table 9 and show that the positional error associated with street geocoding has a comparable effect across unit size as the effect observed when a Local Moran's I technique was used. That is to say, other factors such as crime type, study area, and parameter setting (i.e., cell size) has a much more pronounced effect on predictive accuracy—measured in terms of hit rate, PAI, and RRI—than does diminished positional accuracy of geocoded crime incident locations that are associated with various street geocoded reference data.

Table 9. The influence of cell size[1] and geocoding method on the pedictive accuracy of Getis-Ord Gi* maps of assaults and burglaries, by jurisdiction.

	Assault			Burglary		
	Predictive accuracy			Predictive accuracy		
Jurisdiction, cell size, and geocoding method	Hit rate (%)	PAI	RRI	Hit rate (%)	PAI	RRI
Arlington						
250m						
Address Points	26.34	5.19	0.58	34.44	2.87	0.72
Street Centerlines	25.85	5.10	0.57	35.18	2.93	0.73
NAVTEQ	27.80	5.48	0.61	35.58	2.97	0.74
TeleAtlas	27.32	5.38	0.60	36.03	3.00	0.75
StreetMap	26.34	5.19	0.58	35.35	2.95	0.74
TIGER	25.85	5.10	0.57	35.46	2.96	0.74
Block groups						
Address Points	60.49	1.98	0.72	60.03	1.73	0.91
Street Centerlines	59.51	1.95	0.71	59.98	1.73	0.91
NAVTEQ	60.98	2.00	0.72	60.09	1.74	0.91
TeleAtlas	60.00	1.97	0.71	59.98	1.73	0.91
StreetMap	58.54	1.92	0.69	59.24	1.71	0.90
TIGER	59.02	1.93	0.70	59.29	1.71	0.90
Charlotte						
250m						
Address Points	67.26	8.82	0.94	61.42	5.74	0.88
Street Centerlines	65.09	8.54	0.91	60.27	5.64	0.86
NAVTEQ	66.90	8.78	0.94	61.33	5.74	0.88
TeleAtlas	67.06	8.80	0.94	61.58	5.76	0.88
StreetMap	66.21	8.68	0.93	60.85	5.69	0.87
TIGER	66.84	8.77	0.94	60.88	5.69	0.87

(continued)

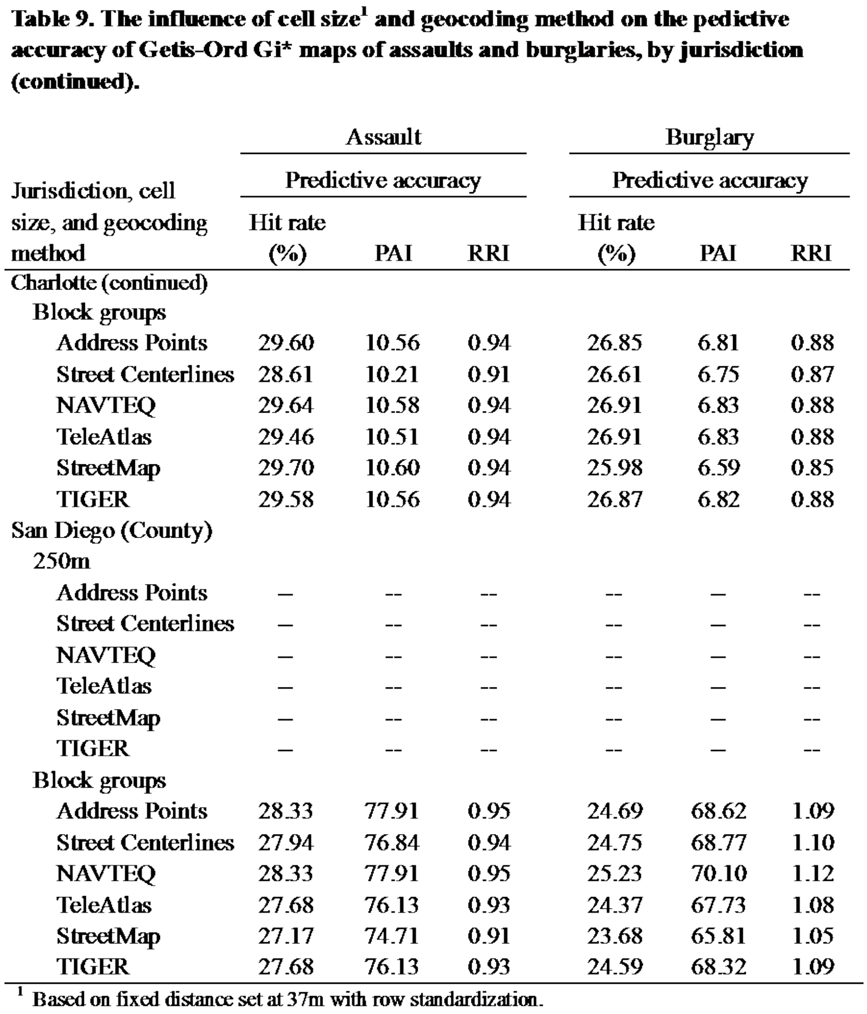


Table 9. The influence of cell size[1] and geocoding method on the pedictive accuracy of Getis-Ord Gi* maps of assaults and burglaries, by jurisdiction (continued).

	Assault			Burglary		
	Predictive accuracy			Predictive accuracy		
Jurisdiction, cell size, and geocoding method	Hit rate (%)	PAI	RRI	Hit rate (%)	PAI	RRI
Charlotte (continued)						
Block groups						
Address Points	29.60	10.56	0.94	26.85	6.81	0.88
Street Centerlines	28.61	10.21	0.91	26.61	6.75	0.87
NAVTEQ	29.64	10.58	0.94	26.91	6.83	0.88
TeleAtlas	29.46	10.51	0.94	26.91	6.83	0.88
StreetMap	29.70	10.60	0.94	25.98	6.59	0.85
TIGER	29.58	10.56	0.94	26.87	6.82	0.88
San Diego (County)						
250m						
Address Points	–	--	--	--	–	--
Street Centerlines	–	--	--	--	–	--
NAVTEQ	–	--	--	--	–	--
TeleAtlas	–	--	--	--	–	--
StreetMap	–	--	--	--	–	--
TIGER	–	--	--	--	–	--
Block groups						
Address Points	28.33	77.91	0.95	24.69	68.62	1.09
Street Centerlines	27.94	76.84	0.94	24.75	68.77	1.10
NAVTEQ	28.33	77.91	0.95	25.23	70.10	1.12
TeleAtlas	27.68	76.13	0.93	24.37	67.73	1.08
StreetMap	27.17	74.71	0.91	23.68	65.81	1.05
TIGER	27.68	76.13	0.93	24.59	68.32	1.09

[1] Based on fixed distance set at 37m with row standardization.

The final hotspot technique we examined is classified as a point-pattern technique. Table 10 provides results from KDE analysis where variations in crime type, study area, parameter setting (i.e., search radius) and geocoding method were considered. Results demonstrate the influence that these factor have on predictive accuracy, gauged by the hit rate, PAI, and RRI.

Findings from the KDE analysis show that reference data quality has a diminished effect on the technique as search radius increases, regardless of study area or crime type considered. For example, the hit rate for assaults in San Diego when KDE uses a 100m-search radius and crime incidents geocoded to address points is 13% versus 3% for crime incidents geocoded to the local street centerline file. However, when the search radius is extended to 500m, the difference in hit rates between the two reference data layers is less than 1% (33.1% versus 33.5%). These findings challenge the growing consensus that KDE produces the most accurate hotspot predictions. Based on our findings, KDE does produce hotspot maps with the highest PAI, but *only* for analysis that 1) uses a small search radius, 2) that rely data geocoded against an *address point* reference layer, and 3) for certain study areas. If KDE uses a larger search radius with data geocoded against lower quality street reference data, for example, then KDE may not be a superior predictive mapping technique.

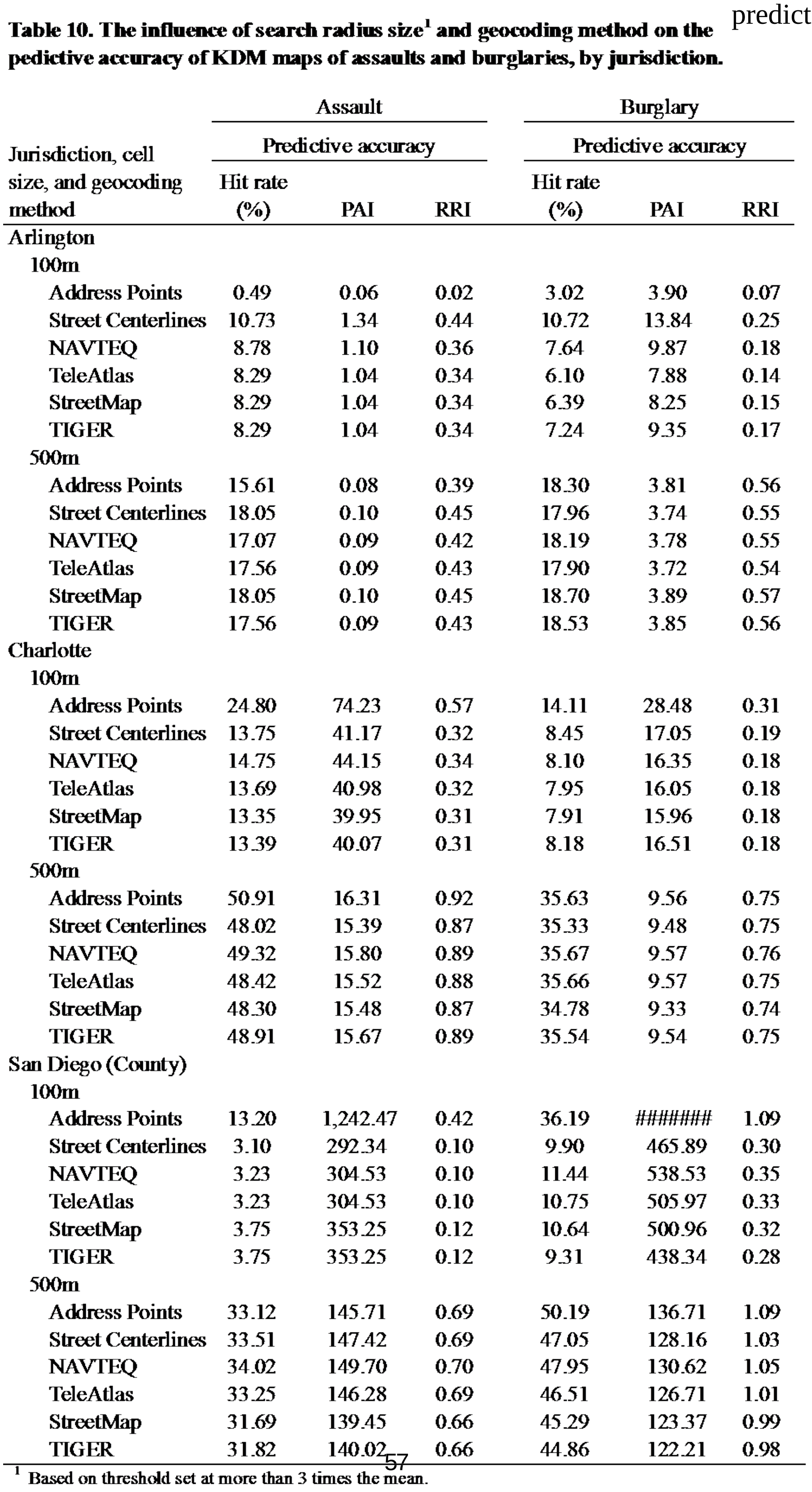


Table 10. The influence of search radius size[1] and geocoding method on the pedictive accuracy of KDM maps of assaults and burglaries, by jurisdiction.

Jurisdiction, cell size, and geocoding method	Assault Predictive accuracy Hit rate (%)	PAI	RRI	Burglary Predictive accuracy Hit rate (%)	PAI	RRI
Arlington						
100m						
Address Points	0.49	0.06	0.02	3.02	3.90	0.07
Street Centerlines	10.73	1.34	0.44	10.72	13.84	0.25
NAVTEQ	8.78	1.10	0.36	7.64	9.87	0.18
TeleAtlas	8.29	1.04	0.34	6.10	7.88	0.14
StreetMap	8.29	1.04	0.34	6.39	8.25	0.15
TIGER	8.29	1.04	0.34	7.24	9.35	0.17
500m						
Address Points	15.61	0.08	0.39	18.30	3.81	0.56
Street Centerlines	18.05	0.10	0.45	17.96	3.74	0.55
NAVTEQ	17.07	0.09	0.42	18.19	3.78	0.55
TeleAtlas	17.56	0.09	0.43	17.90	3.72	0.54
StreetMap	18.05	0.10	0.45	18.70	3.89	0.57
TIGER	17.56	0.09	0.43	18.53	3.85	0.56
Charlotte						
100m						
Address Points	24.80	74.23	0.57	14.11	28.48	0.31
Street Centerlines	13.75	41.17	0.32	8.45	17.05	0.19
NAVTEQ	14.75	44.15	0.34	8.10	16.35	0.18
TeleAtlas	13.69	40.98	0.32	7.95	16.05	0.18
StreetMap	13.35	39.95	0.31	7.91	15.96	0.18
TIGER	13.39	40.07	0.31	8.18	16.51	0.18
500m						
Address Points	50.91	16.31	0.92	35.63	9.56	0.75
Street Centerlines	48.02	15.39	0.87	35.33	9.48	0.75
NAVTEQ	49.32	15.80	0.89	35.67	9.57	0.76
TeleAtlas	48.42	15.52	0.88	35.66	9.57	0.75
StreetMap	48.30	15.48	0.87	34.78	9.33	0.74
TIGER	48.91	15.67	0.89	35.54	9.54	0.75
San Diego (County)						
100m						
Address Points	13.20	1,242.47	0.42	36.19	#######	1.09
Street Centerlines	3.10	292.34	0.10	9.90	465.89	0.30
NAVTEQ	3.23	304.53	0.10	11.44	538.53	0.35
TeleAtlas	3.23	304.53	0.10	10.75	505.97	0.33
StreetMap	3.75	353.25	0.12	10.64	500.96	0.32
TIGER	3.75	353.25	0.12	9.31	438.34	0.28
500m						
Address Points	33.12	145.71	0.69	50.19	136.71	1.09
Street Centerlines	33.51	147.42	0.69	47.05	128.16	1.03
NAVTEQ	34.02	149.70	0.70	47.95	130.62	1.05
TeleAtlas	33.25	146.28	0.69	46.51	126.71	1.01
StreetMap	31.69	139.45	0.66	45.29	123.37	0.99
TIGER	31.82	140.02	0.66	44.86	122.21	0.98

[1] Based on threshold set at more than 3 times the mean.

5.4. Discussion and Conclusions

The second component of our study provides answers to important questions related to the predictive accuracy of hotspot mapping. First, results of our analysis reveal that several factors affect the reliability of predictive hotspot mapping. For example, findings suggest that hotspot technique substantially affects predictive accuracy (see Table 4). In terms of a hit rate, the least accurate technique in terms of predicting future crime events was grid-based thematic mapping (7%), compared to the most accurate, which was a Local Moran's I analysis that aggregated crime data to block groups (47%). According to an alternative measure of accuracy (i.e., the PAI), however, a Gi* approach that aggregated incident locations to block groups was least accurate (PAI=14) and STAC analysis was most accurate (PAI=456). Finally, when the size of the study area was factored into the metric used to assess predictive accuracy, far less variation was observed across technique. Still, grid-based thematic mapping was shown to be the least accurate (RRI=.69) and a grid-based Gi* approach was the most accurate.

Results of our analysis also indicate that crime type has a moderate effect on predictive accuracy for every technique considered. For example, homicide events consistently have the lowest hit rates, regardless of mapping technique used. A hit rate of less than 1% was observed for homicides mapped using KDE, less than 2% using grid-based thematic mapping, and less than 4% using Local Moran's I (grid-based) (see Table 5). On the other hand, regardless of technique, drug offenses was the crime type often associated with the highest hit rates. The crime types associated with the highest and lowest hit rates, however, did not also correspond to the highest and lowest PAI and RRI. For example, the difference in PAI scores between homicides (PAI=71) and robberies (PAI=72) that were mapped using a grid-based thematic mapping technique were much more comparable than their corresponding RRI scores (i.e., .07 and .72, respectively). Again, these findings suggest that crime type influences the predictive accuracy of hotspot mapping.

Third, our findings also indicate that study area has a modest effect on predictive accuracy, and that its effect varies by mapping technique and metric used to determine accuracy (see Table 6). For example, Las Vegas produced the highest hit rates and RRI scores when grid-based thematic mapping (hit rate=44%; RRI=.25), KDE (hit rate=38%; RRI=.19), NNH (hit rate=20%; RRI=.05), and STAC (hit rate=5%; RRI=0.004) techniques were used. However, the highest PAI scores for these techniques were associated with San Diego County (PAI=191, PAI=329, PAI=2,158, and PAI=2,614, respectively).

Fourth, based on our analysis of prediction maps produced using KDE, regardless of crime type, results showed that the effects of parameter settings are substantial (see Figures 11-17). However, the impact of these effects varies by predictive metric used to assess accuracy. For example, neither the effects of kernel bandwidth nor cell size appear to have a pronounced effect on the RRI. And while there is a corresponding increase in hit rate to an increase in kernel bandwidth, the increase is not proportionate to the increase in the search radius. For example, doubling the bandwidth from 50m to 100m or from 100m to 200m does not result in a corresponding two-fold increase in hit rate.

Finally, the effect of geocoding quality on predictive hotspot mapping appears to be mitigated by all of factors considered and described above. For example, analysis of grid-based thematic mapping techniques (see Table 7) suggest that the effects of geocoding quality of predictive hotspot crime mapping are mitigated by crime type, study area, and cell size. For other types of aggregate techniques such as Local Moran's I and Gi* (see Tables 8 and 9, respectively), the effects of geocoding quality on predictive hotspot crime mapping is less apparent and does not change considerably when crime type, study area, and cell size are taken into consideration. When point-pattern analysis such as KDE was examined, however, the effects of geocoding quality on predictive hotspot mapping seemed to be influenced by some factors, but not by others (see Table 10). For example, the impact of geocoding quality varied considerably by crime type and urban morphology; but became inconsequential as a key parameter setting (i.e., the search radius) was increased. Figure 18 illustrates how this dynamic works and shows how variations in two key parameter settings used in KDE mapping (i.e., search radius and cell size threshold) will influence what is defined as a hotspot, which in turn will affect measures of predictive accuracy.

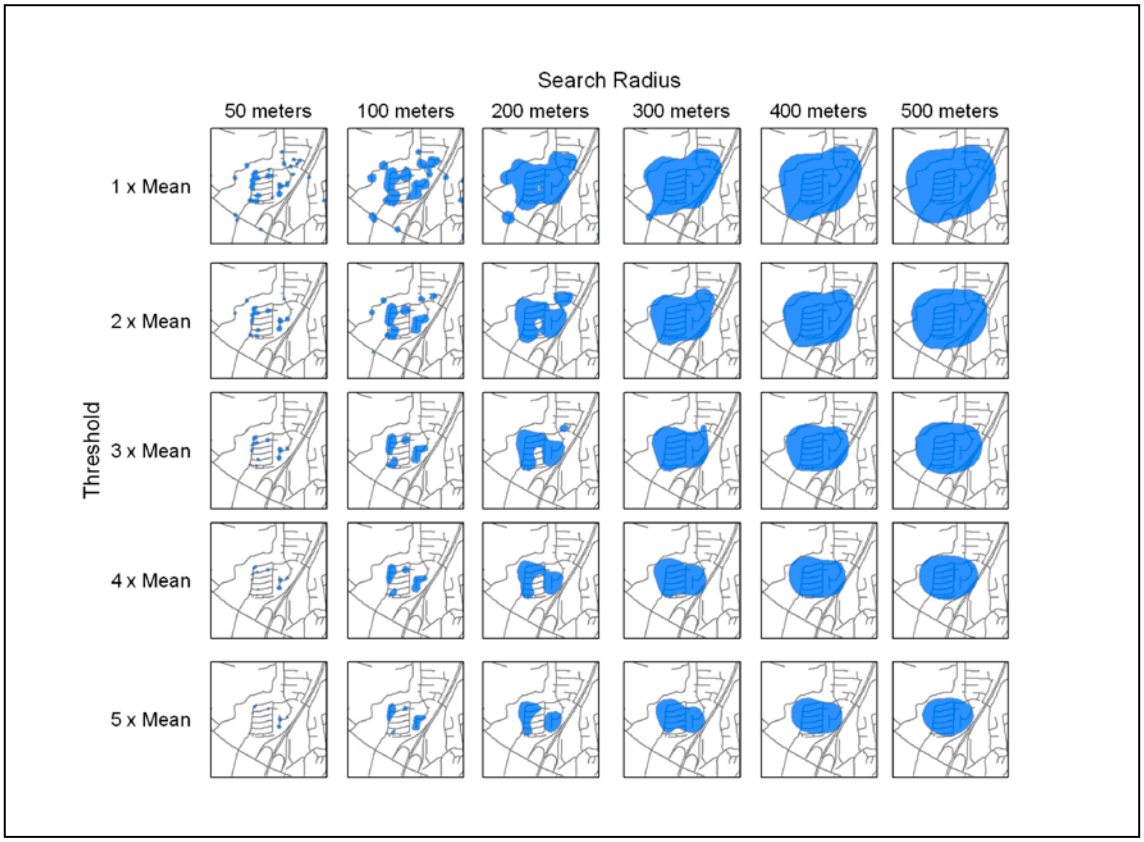


Figure 18. Variations in two key parameter settings used to produce KDE maps. As search radius size increases, the effects of geocoding quality on predictive hotspot mapping decreases. However, the loss of precision can result. Similarly, a corresponding increase in the cell size threshold can also adversely impact prediction results.

In summary, results of our analysis suggest that no one predictive hotspot mapping technique is superior to any other. This is due, in part, to how factors such as crime type, urban morphology, and parameter settings can impact the predictive accuracy of hotspot techniques differently. Moreover, the effects of geocoding quality on predictive hotspot crime mapping is also varies by technique as well as by many of the other factors considered. Nevertheless, recommendations for future research in this area can be offered as well as recommendations for analysts seeking the optimal approach to analyzing crime patters. Before these recommendations are offered, however, a discussion of the study's limitations is offered.

5.5. Limitations

Many of the issues that limit the substantive conclusions produced from our analysis of geocoding quality (see Section 4.5) also limit conclusions reached with respect to our analysis of predictive hotspot mapping. Two specific limitations are noteworthy, however. First, as noted previously, the data used in the second component of our study is limited to six geographic areas. As a result, patterns observed with respect to the predictive accuracy of various hotspot techniques considered may not be applicable to other agencies.

And second, although we considered a number of factors that could influence the predictive accuracy of hotspot analysis, which in turn influences the effects of geocoding quality, not every factor—or iteration of factors combined—was analyzed. Granted, we examined a number of combinations of factors that we felt would not only help us better understand the role of geocoding quality on predictive hotspot mapping, but that also reflected decision-making points the analysis process (i.e., should crime events be disaggregated, what hotspot method is most appropriate, how does the urban morphology affect results, and what are the optimal parameter settings for the technique that is being used, etc.). Still, considering every possible combination of factors that play a role in the outcome of hotspot analysis simply is not possible; and therefore prevents us from pinpointing a single "best technique" for analysts and academics to use when conducting crime-pattern analysis.

5.6. Future Research

Although the limitations described in the previous section may restrict our inferences about the effects of geocoding quality on predictive hotspot mapping, these problems do not limit the importance of current findings, the study's contribution to our broader understanding of these techniques as they apply to the analysis of crime events, or the implications that our findings have on future research. We feel that two areas in particular warrant immediate investigation.

First, future research should explore how measures of predictive accuracy can be enhanced. Each of the three metrics used in the current study (i.e., hit rate, PAI, and RRI) are relative measures and objective thresholds that define what is a "good" predictive hotspot map versus what constitutes a "bad" prediction do not exist. For example, a grid-based analysis of burglaries in San Diego (County), using a 250m grid-cell size, and relying on data geocoded against Census TIGER reference data produced a hit rate of slightly less than 3%. Comparatively, the hit rate for hotspot analysis of burglary events in Arlington consistently produced a hit rate around 82%, regardless of reference data used. However, any assessment of a “good” hit rate is merely subjective. Furthermore, comparisons across measures of predictive accuracy should be made with caution, since factors such as the size of the study area can substantially influence results and are not based on the same scale. That is, the hit rate is restricted to a score that ranges between 0% and 100%, whereas the PAI and the RRI scores are not.

Second, future research should explore ways to enhance traditional output of hotspot analysis. In our view, researchers and analysts would benefit from results of predictive hotspot mapping that illustrated how variations in different assumptions associated with any given technique affect outcomes. For example, Figure 18 illustrates how changing two key parameter settings used to produce KDE maps influence prediction results. Specifically, as search radius size increases, the effects of geocoding quality on predictive hotspot mapping decreases. However, the loss of precision can result. Similarly, a corresponding increase in the cell size threshold can also adversely impact findings. Output such that shown in Figure 18 would go far in helping researchers and practitioners make an informed decision about parameter setting associated with any prediction technique. Therefore, future research should focus on ways hotspot analysis output can be enhanced.

5.7. Recommendations

Collectively, results from the current study suggest that several factors, including hotspot technique, crime type, and study area affect predictive hotspot mapping. Results also demonstrated that prediction metrics are sensitive to user-defined parameter settings and geocoding quality. These findings lead us to conclude that no single hotspot technique is more accurate than any other. Instead, they are highly influenced by these factors. Nevertheless, a series of recommendations based on our results are presented below and offer guidance for analysts and researchers engaging in hotspot analysis.

1. **Consider analyzing data with multiple techniques**

   Findings show that predictive accuracy varies substantially across the six different hotspot techniques examined in the current study. Since these approaches vary in their assumption and limitations, and rely on one or more user-defined parameters, analysts and researchers should consider analyzing their data using multiple methods.

2. **Disaggregate crime incidents and analyze like crime events separately**

   Findings show that crime type has some effect on predictive accuracy of hotspot mapping techniques. For example, predictive accuracy is considerably lower for homicide than for other crime types, which is likely due to the relatively fewer number of homicides than other types of crime. Therefore, it is recommended that hotspot analysis disaggregate crime incidents and analyze like crime events separately.

3. Take study area into consideration

Current findings show that urban morphology has a modest effect on predictive accuracy. Depending on the metric used to measure accuracy and the size of the study area, however, this effect can be very pronounced. Therefore, consideration should be given to urban morphology prior to conducting hotspot analysis.

4. Be cognizant of user-defined parameter settings

Hotspot analysis techniques examined in the current study vary in their assumption and limitations. More importantly, they all rely on one or more user-defined parameters. Hotspot techniques are sensitive to the selection of these parameters and some of the disagreement over which techniques performs best (e.g., Chainey et al., 2008 versus Levine, 2008) may in fact partly be due to differences in parameter selection. Therefore, researchers and analysts must be cognizant of user-defined parameter settings and how modifications to these settings may affect prediction maps.

5. Use street centerline reference data or address point reference data

The effect of geocoding quality on predictive hotspot mapping is complex and varies depending on prediction technique (and parameter settings within them), crime type, and urban morphology. Address point and street centerline reference data were frequently associated with the best prediction results. Therefore, it is recommended that to limit the effects of geocoding quality on predictive hotspot mapping that researchers and analysts rely on data that have been geocoded to either address point or street centerline reference files.

6. Determine how predicative accuracy will be measured

Finally, we used three measures of predictive accuracy in the current study: 1) hit rate; 2) Predictive Accuracy Index (PAI) (Chainey et al., 2008); and 3) Recapture Rate Index (RRI) (Levine, 2008). When variations in hotspot technique, crime type, study area, parameter settings, and geocoding quality were considered, these metrics often produced inconsistent results. Some of these inconsistencies can be explained by how the metrics are computed. For example, since the PAI uses the size of the hotspots (i.e., the hotspot area) in its calculation, the difference between a PAI score and a RRI score often depended on the size of the study area. Therefore, prior to conducting hotspot analysis, a determination should be made on how predictive accuracy will be measured. This decision should be based, in part, on the size of the study area.

6. REFERENCES

Anderson, T. (2006). Comparison of spatial methods for measuring road accident 'hotspots': A case study of London. *Journal of Maps, 2006*, 55-63.

Andresen, M.A. (2006). Crime measures and the spatial analysis of criminal activity. *British Journal of Criminology, 46*, 258-285.

Anselin, L., Cohen, J. Cook, D., Gorr, W. & Tita, G. (2000). Spatial analyses of crime. *Criminal Justice, 4*, 213-262.

Bernasco, W., & Nieuwbeerta, P. (2005). How do residential burglars select target areas? *British Journal of Criminology, 44*, 296-315.

Bichler, G., & Balchak, S. (2007). Address matching bias: Ignorance is not bliss. *Policing: An International Journal of Police Strategies and Management*, *30*(1), 32-60.

Boba-Santos, R.L. (2012). *Crime analysis and crime mapping (3rd ed)*. Sage: Thousand Oaks, CA.

Bonner, M.R., Han, D., Nie, J., Rogerson, P., Vena, J.E., & Freudenheim, J.L. (2003). Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology, 14*(4), 408-412.

Bowers, K.J., & Johnson, S.D. (2003). Measuring the geographic displacement and diffusion of benefits of crime prevention activity. *Journal of Quantitative Criminology, 19*(3), 275-301.

Bowers, K.J., Johnson, S.D., & Pease, K. (2004). Prospective hotspotting. The future of crime mapping? *British Journal of Criminology, 44*, 641-658.

Brimicombe, A.J., Brimicombe, L.C., & Li., Y. (2007). Improving geocoding rates in preparation for crime data analysis. *International Journal of Police Science and Management, 9*(1), 80-92.

Brimicombe, A.J. (2005). Cluster detection in point event data having tendency towards spatially repetitive events. GeoComputation 2005. 8th International Conference of GeoComputation, August 1-3, Ann Arbor, Michigan.

Britt, H.R., Carlin, B.P., Toomey, T.L., & Wagenaar, A.C. (2005). Neighborhood level spatial analysis of the relationship between alcohol outlet density and criminal violence. *Environmental and Ecological Statistics, 12*, 411-426.

Burra, T., Jerrett, M., Burnett, R.T., & Anderson, M. (2002). Conceptual and practical issues in the detection of local disease clusters: A study of mortality in Hamilton, Ontario. *The Canadian Geographer, 46*, 160-171.

Cahill, M. & Mulligan, G. (2007). Using geographically weighted regression to explore local crime patterns. *Social Science Computer Review, 25*(2), 174-193.

Cayo, M.R., & Talbot, T.O. (2003). Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics, 2*(10), 1-12.

Chainey, S. (2005). How accurate is my hotspot map? Eighth Annual Crime Mapping Research Conference, September 7-10, Savannah, GA.

Chainey, S., & Ratcliffe, J.H. (2005). *GIS and crime mapping*. Wiley: San Francisco, CA.

Chainey, S., & Thompson, L. (Eds.) (2008a). *Crime mapping case-studies: Practice and research*. San Francisco, CA: Wiley.

Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal 21*, 4-28.

Christen, P., Churches, T., & Zhu, J.X. (2002). Probabilistic name and address cleaning and standardization. The Australasian Data Mining Conference.

Christens, B., & Speer, P.W. (2005). Predicting violent crime using urban and suburban densities. *Behavior and Social Studies, 14*, 113-127.

Dearwent, S.M., Jacobs, R.J., & Halbert, J.B. (2001). Locational uncertainty in georeferencing public health datasets. *Journal of Exposure Analysis and Environmental Epidemiology, 11*, 329-334.

Doran, B.J., & Lees, B.G. (2005). Investigating the spatiotemporal links between disorder, crime and fear of crime. *The Professional Geographer, 57*(1), 1-12.

Eck, J.E., Chainey, S., Cameron, J.G., Leitner, M., & Wilson, R.E. (2005). *Mapping crime: Understanding hotspots*. National Institute of Justice Special Report.

Gilboa, S.M., Mendola, P., Olshan, A.F., Harness, C., Loomis, D., Langlois, P.H., Savitz, D.A., & Herring, A.H. (2006). Comparison of residential geocoding methods in population-based study of air quality and birth defects. *Environmental Research, 101*, 256-262.

Goldberg, D.W., Wilson, J.P., & Knoclock, C.A. (2007). From text to geographic coordinates: The current state of geocoding. *URISA journal*, *19*(1), 33-46.

Groff, E.R., & LaVigne, N.G. (2002). Forecasting the future of predictive crime mapping. *Crime Prevention Studies, 13*, 29-57.

Groff, E.R., & LaVigne, N.G. (2001). Mapping an opportunity surface of residential burglary. *Journal of Research in Crime and Delinquency, 38*(3), 257-278.

Grubesic, T.H. (2006). On the application of fuzzy clustering for crime hot spot detection. *Journal of Quantitative Criminology, 22*(1), 77-105.

Grubesic, T.H., Mack, E., & Murray, A.T. (2007). Geographic exclusion: Spatial analysis for evaluating the implications of Megan's Law. *Social Science Computer Review, 25*(2), 143-162.

Harada, Y., & Shimada, T. (2006). Examining the impact of the precision of address geocoding on estimated density of crime locations. *Computers & Geoscience, 32*, 1096-1107.

Harries, K. (1999). *Mapping crime: Principle and Practice*. United States Department of Justice, National Institute of Justice.

Hirschfield, A. (2007). *Mapping and analyzing crime data: Lessons from research and practice*. Taylor and Francis.

Jacquez, G.M., & Waller, L. (2000). The effect of uncertain locations on disease cluster statistics. In H.T. Mowrer & R.G. Congalton (Eds.), *Quantifying spatial uncertainty in natural resources: Theory and applications for GIS and Remote Sensing (pp. 53-64)*. Chelsea, Michigan: Arbor Press.

Johnson, S.D., Browers, K.J., Birks, D.J., & Pease, K. (2008). Predictive mapping of crime by ProMap: Accuracy, units of analysis and the environmental backcloth. In D. Weisburd, W. Bernasco and G.J.N. Bruinsma (eds.). *Putting crime in its place: Units of analysis in geographic criminology (Chapter 8, pp. 171-198)*. Springer, New York.

Johnson, S.D., Lab, S.P., & Browers, K.J. (2008). Stable and fluid hotspots of crime: Differentiation and identification. *Built Environment, 34*(1), 32-45.

Johnson, S.D., & Bowers, K.J. (2004). The burglary as clue to the future. The beginnings of prospective hot-spotting. *European Journal of Criminology, 1*(2), 237-255.

Karimi, H.A., & Durcik, M. (2004). Evaluation of uncertainties associated with geocoding techniques. *Computer-Aided Civil and Infrastructure Engineering, 19*, 170-185.

LaVigne, N.G. (2007). *Mapping for community-based prisoner reentry efforts: A guidebook for law enforcement agencies and their partners*. Policy Foundation and the Office of Community Oriented Policing Services.

LaGrange, T.L. (1999). The impact of neighborhoods, schools and malls on the spatial distribution of property damage. *Journal of Research in Crime and Delinquency, 36*(4), 393-421.

Lersch, K.M., & Hart, T.C. (2011). *Space, Time, and Crime (3rd ed.)*. Durham, NC: Carolina Press.

Levine, N. (2008). The "hottest" part of a hotspot: Comment on "The utility of hotspot mapping for predicting spatial patterns of crime". *Security Journal, 21*, 295-302

Levine, N., & Kim, K.E. (1998). The location of motor vehicle crashes in Honolulu: A methodology for geocoding intersections. *Computers, Environment and Urban Systems, 22*(6), 557-576.

McLafferty, S., Williamson, D., & Maguire, P.G. (2000). Identifying crime hot spots using kernel smoothing'. In V. Goldsmith, et al. (Eds.), *Analyzing crime patterns: Frontiers of practice (pp. 77-85)*. Thousand Oaks, CA: Sage.

Murray, A.T., McGuffog, I., Western, J.S., & Mullins, P. (2001). Exploratory spatial data analysis techniques for examining urban crime. *British Journal of Criminology, 41*, 309-329.

Oliver, M.N., Matthews, K.A., Siadaty, M., Hauck, F.R., & Pickle, L.W. (2005). Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics 4*(29).

Paulsen, D.J., & Robinson, M.B. (2008). *Crime mapping and spatial aspects of crime (2nd ed.)*. Allyn & Bacon: Boston, MA.

Poulsen, E., & Kennedy, L.W. (2004). Using dasymetric mapping for spatially aggregated crime data. *Journal of Quantitative Criminology, 20*(3), 243-262.

Ratcliffe, J.H. (2005). Detecting spatial movement of intra-region crime patterns over time. *Journal of Quantitative Criminology, 21*(1), 103-123.

Ratcliffe, J.H. (2004). Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science, 18*(1), 61-72.

Ratcliffe, J.H. (2001). On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. *International Journal of Geographical Information Science, 15*(5), 473-485.

Ratcliffe, J.H., & McCullagh, M.J. (2001). Chasing ghosts? Police perception of high crime areas. *British Journal of Criminology, 41*, 330-341.

Ratcliffe, J.H., & McCullagh, M.J. (1998). Identifying repeat victimization with GIS. *British Journal of Criminology, 38(4)*, 651-662.

Reaves, B.A., & Hart, T.C. (2000). *Law Enforcement and Management and Administrative Statistics, 1999: Data for Individual State and Local Agencies with 100 or More Officers*. Bureau of Justice Statistics. Washington, DC: Government Printing Office. NCJ 184481.

Rushton, G., Armstrong, M.P., Gittler, J., Greene, B., Pavlik, C.E., West, M.W., & Zimmerman, D.L. (2006). Geocoding in cancer research: A review. *American Journal of Preventative*

Medicine, 30(2S), S16-S24.

Schootman, M., Sterling, D.A., Struthers, J., Yan, Y., Laboube, T., Emo, B., & Higgs, G. (2007). Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Annals of Epidemiology, 17*(6), 464-470.

Strickland, M.J., Siffel, C., Gardner, B.R., Berzen, A.K., & Correa, A. (2007). Quantifying geocode location error using GIS methods. *Environmental Health, 6*(10).

Tompson, L., Partridge, H., & Shepherd, N. (2009). Hot routes: Developing a new technique of the spatial analysis of crime. *Crime Mapping: A Journal of Research and Practice, 1*(1), 77-96.

Waller, L.A. (1996). Statistical power and design of focused clustering studies. *Statistics in Medicine, 15*, 765-782.

Wang, F. (2005a). Job access and homicide patterns in Chicago: An analysis at multiple geographic levels based on scale-space theory. *Journal of Quantitative Criminology, 21*(2), 195-217.

Wang, F. (2005b). *Geographic information systems and crime analysis*. IGI Global.

Ward, M.H., Nuckols, J.R., Giglierano, J., Bonner, M.R., Wolter, C., Airola, M., Mix, W., Colt, J., & Hartge, P. (2005). Positional accuracy of two methods of geocoding. *Epidemiology, 16*(4), 542-547.

Whitsel, E.A., Rose, K.M., Wood, J.L., Henley, A.C., Liao, D., & Heiss, G. (2004). Accuracy and repeatability of commercial geocoding. *American Journal of Epidemiology, 160*(10), 1023-1029.

Whitsel, E.A., Quibrera, P.M., Smith, R.L., Catellier, D.J., Liao, D., Henley, A.C., & Heiss, G. (2006). Accuracy of commercial geocoding: Assessment and implications. *Epidemiological Perspectives and Innovations, 3*(8), 1-12.

Zandbergen, P.A. (2007). Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health, 7*, 37.

Zandbergen, P.A. (2008a). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems, 32*(3), 214-232.

Zandbergen, P.A. (2008b). Positional accuracy of spatial data: Non-normal distributions and a critique of the National Standard for Spatial Data Accuracy. *Transactions in GIS, 12*(1), 103-130.

Zandbergen, P.A. (2009). Geocoding quality and implications for spatial analysis. *Geography Compass, 3*(2), 647-680.

Zandbergen, P.A., & Green, J.W. (2007). Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. *Environmental Health Perspectives, 115*(9), 1363-1370.

Zandbergen, P.A., & Hart, T.C. (2006). Reducing housing options for convicted sex offenders: Investigating the impact of residency restriction laws using GIS. *Justice Research and Policy, 8*, 1-24.

Zandbergen, P.A., & Hart, T.C. (2009). Geocoding accuracy considerations in determining residency restrictions for sex offenders. *Criminal Justice Policy Review, 20*(1), 62-90.

Zhan, F.B., Brender, J.D., De Lima, I., Suarez, L., & Langlois, P.H. (2006). Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Annals of Epidemiology, 16*(11), 842-849.

Zimmerman, D.L., Fang, X., Mazumdar, S., & Rushton, G. (2007). Modeling the probability distribution of positional errors incurred by residential address geocoding. International *Journal of Health Geographics*, *6*(1).

Zimmerman, D.L. (2008). Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics*, *64*(1), 262–270.

7. DISSEMINATION OF INFORMATION

Our dissemination strategy focused primarily on presentations of findings at conferences, submitting our findings to scholarly journals for publication, leading relevant workshops, and submitting a final report to NIJ and (with permission of NIJ) the distribution of a condensed version of the final report to interested parties, including crime mapping listservs and user groups. Summaries of each of these efforts are described below in greater detail.

7.1. Conference Presentations

Zandbergen, P.A., Hart, T.C., & Camponovo, M.E. (2012). *Predictive crime hotspot mapping*. Annual meeting of the Association of American Geographers. February 24-28, 2012. New York City, NY.

Zandbergen, P.A., Hart, T.C., Lenzer, K.E., & Camponovo, M.E. (2011). *Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets*. 1st International Geocoding Conference, December 6-7, 2011, Redlands, CA.

Camponovo, M.E., & Zandbergen, P.A. (2011). *Robustness of kernel density crime hotspot maps*. ESRI Southwest User Conference, November 16-18, 2011, Mesa, AZ.

Hart, T.C., & Zandbergen, P.A. (2011). *Effects of geocoding quality on predictive hotspot mapping*. American Society of Criminology Annual Meeting, November 16-19, 2011, Washington, DC.

Hart, T.C., & Zandbergen, P.A. (2011). *Effects of geocoding quality on predictive hotspot mapping*. 11th Crime Mapping Research Conference, October 19-21, 2011, Miami, FL.

Zandbergen, P.A. (2011). *Predictive crime hotspot mapping*. ESRI International User Conference, July 11-15, 2011, San Diego, CA.

Hart, T.C., & Zandbergen, P.A. (2010). *Geocoding crime incident locations: Assessing match rates and positional accuracy across techniques*. American Society of Criminology Annual Meeting, November 17-20, 2010, San Francisco, CA.

Zandbergen, P.A. & Hart, T.C. (2010). *Effects of data quality on predictive hotspot mapping*. National Institute of Justice Conference, June 14-16, 2010, Arlington, VA.

7.2. Journal Article Submissions/Publications

Hart, T.C., & Zandbergen, P.A. (In Press—*Policing: An International Journal of Police Strategies & Management*). Reference data and geocoding quality: Examining completeness and positional accuracy of street geocoded crime incidents.

Zandbergen, P.A., Hart, T.C., Lenzer, K.E., & Camponovo, M.E. (2012). Error propagation models to examine the effects of geocoding quality of spatial analysis of individual-level datasets. *Spatial and Spatio-temporal Epidemiology, 3*(1), 69-82.

7.3. Workshops

Hart, T.C., (2012). *Geocoding Crime Data*. Workshop to be conducted at the International Association of Crime Analysts (IACA) Conference to be held in Henderson, NV; September, 2012.

Hart, T.C. (2011). *Geocoding crime data*. Workshop conducted at the 11th Crime Mapping Research Conference, October 19-21, 2011, Miami, FL.

Zandbergen, P.A. (2011). *Python scripting for the automation of GIS Workflows, Part I*. Workshop conducted at the 11th NIJ Crime Mapping Conference, October 19-21, 2011, Miami, FL.

Zandbergen, P.A. (2011). *Python scripting for the automation of GIS Workflows, Part II*. Workshop conducted at the 11th NIJ Crime Mapping Conference, October 19-21, 2011, Miami, FL.

Zandbergen, P.A. (2011). *Crime hotspot mapping and analysis*. Workshop conducted at the 11th NIJ Crime Mapping Conference, October 19-21, 2011, Miami, FL.

Appendix A

Table A1. Geocoding match rate results for Albuquerque Police Department by crime type and type of street reference data, 2007-08.

		Street Geocoding				
		Free			Commercial	
Crime events in --	Average	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas
Albuquerque						
All locations	84.1	87.0	77.3	75.8	89.0	91.4
Without intersections	84.2	87.2	77.4	75.9	89.2	91.5
Assaults	84.1	87.1	76.7	74.5	89.6	92.7
Auto burglary	83.4	88.1	76.4	75.0	87.3	90.3
Auto theft	84.0	85.9	78.0	76.7	88.9	90.5
Burglary	85.1	89.0	76.3	76.7	91.4	92.1
Drug offenses	85.8	81.6	83.2	79.3	91.1	93.7
Homicide	80.9	79.1	77.9	68.6	88.4	90.7
Robbery	86.0	83.1	83.1	77.3	91.4	95.2
Intersections only	66.8	71.2	68.2	63.2	60.2	71.2
Assaults	68.3	75.4	72.0	62.9	59.4	72.0
Auto burglary	71.0	72.8	71.2	68.0	68.8	74.4
Auto theft	61.0	62.7	55.4	59.0	61.4	66.3
Burglary	52.2	52.2	47.8	52.2	47.8	60.9
Drug offenses	50.3	56.3	51.6	45.3	43.8	54.7
Homicide	--	—	—	—	--	--
Robbery	84.2	90.3	93.5	82.3	64.5	90.3

Note: Drug offenses include drug possession, trafficking, and narcotics incidents known to law enforcement.

— No intersection only addresses.

Table A2. Geocoding match rate results for Arlington (TX) Police Department by crime type and type of street reference data, 2007-08.

		Street Geocoding				
		Free			Commercial	
Crime events in --	Average	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas
Arlington (TX)						
All locations	89.7	96.4	83.4	79.0	94.8	94.9
Without intersections	89.7	96.5	83.4	79.0	94.8	94.9
Assaults	90.1	97.2	83.0	78.9	95.7	96.0
Auto burglary	88.9	95.9	83.8	77.4	93.5	93.9
Auto theft	87.7	95.1	79.9	76.2	93.6	93.6
Burglary	92.4	98.1	84.7	84.8	97.4	97.0
Drug offenses	86.9	96.2	75.8	68.5	96.7	96.9
Homicide	85.6	94.4	69.4	69.4	97.2	97.2
Robbery	91.1	98.0	84.7	78.3	97.5	96.9
Intersections only	33.0	0.0	26.1	52.2	43.5	43.5
Assaults	0.0	0.0	0.0	0.0	0.0	0.0
Auto burglary	27.3	0.0	9.1	54.5	36.4	36.4
Auto theft	40.0	0.0	60.0	60.0	40.0	40.0
Burglary	--	—	—	—	--	--
Drug offenses	0.0	0.0	0.0	0.0	0.0	0.0
Homicide	--	—	—	—	--	--
Robbery	52.0	0.0	40.0	60.0	80.0	80.0

Note: Drug offenses include drug possession, trafficking, and narcotics incidents known to law enforcement.

— No intersection only addresses.

Table A3. Geocoding match rate results for Charlotte-Mecklenburg Police Department by crime type and type of street reference data, 2007-08.

		Street Geocoding				
		Free			Commercial	
Crime events in --	Average	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas
Charlotte-Mecklenburg						
All locations	80.6	92.1	70.6	70.5	84.9	84.6
Without intersections	84.0	97.0	73.6	72.8	88.4	88.0
Assaults	85.7	96.4	77.4	76.1	89.2	89.4
Auto burglary	81.4	97.2	69.4	67.8	86.9	85.6
Auto theft	83.4	96.3	72.4	73.4	87.6	87.4
Burglary	86.3	98.1	73.6	77.7	91.9	90.1
Drug offenses	83.7	95.6	79.2	70.5	85.5	87.7
Homicide	85.9	97.7	77.3	75.0	92.2	87.5
Robbery	84.0	97.3	78.2	70.7	85.4	88.3
Intersections only	18.3	2.3	16.8	28.5	20.0	23.8
Assaults	17.9	2.1	16.0	27.8	19.8	23.6
Auto burglary	20.0	2.2	21.4	29.7	22.5	24.2
Auto theft	23.7	3.6	23.4	35.5	24.9	31.0
Burglary	17.6	2.1	16.6	28.5	18.8	21.8
Drug offenses	8.0	13.3	0.0	26.7	0.0	0.0
Homicide	23.5	0.0	26.1	34.8	26.1	30.4
Robbery	21.2	3.0	21.3	32.1	22.3	27.5

Note: Drug offenses include drug possession, trafficking, and narcotics incidents known to law enforcement.

Table A4. Geocoding match rate results for Las Vegas Metropolitan Police Department by crime type and type of street reference data, 2007-08.

		Street Geocoding				
		Free			Commercial	
Crime events in --	Average	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas
Las Vegas						
All locations	77.2	84.9	69.2	65.9	82.4	83.8
Without intersections	84.2	92.8	74.8	71.1	90.7	91.7
Assaults	85.0	92.1	77.6	72.5	90.9	92.1
Auto burglary	81.7	91.0	70.7	68.3	88.6	90.0
Auto theft	85.8	93.6	77.6	72.2	91.9	93.4
Burglary	82.9	92.4	70.9	70.4	90.5	90.2
Drug offenses	87.1	95.0	81.3	73.2	92.1	93.7
Homicide	83.1	88.8	73.5	72.6	87.4	93.3
Robbery	86.6	92.9	82.0	73.6	91.1	93.5
Intersections only	35.7	36.0	38.0	35.0	32.9	36.8
Assaults	39.9	43.7	42.2	39.8	32.1	41.9
Auto burglary	25.7	28.4	26.6	26.6	18.7	28.3
Auto theft	41.7	45.1	41.9	41.4	35.3	44.7
Burglary	16.5	19.8	14.5	15.1	16.9	16.0
Drug offenses	34.1	36.0	34.3	33.0	32.9	34.4
Homicide	12.2	13.9	16.7	13.9	8.3	8.3
Robbery	49.1	51.7	48.9	48.3	45.1	51.2

Note: Drug offenses include drug possession, trafficking, and narcotics incidents known to law enforcement.

Table A5. Geocoding match rate results for San Diego (CA) County Sheriff's Office by crime type and type of street reference data, 2007-08.

		Street Geocoding				
		Free			Commercial	
Crime events in --	Average	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas
San Diego County						
All locations	78.8	87.9	83.9	79.2	75.0	67.8
Without intersections	82.6	91.4	87.1	81.8	81.1	71.4
Assaults	85.7	92.0	87.9	82.8	88.4	77.5
Auto burglary	84.1	91.2	86.5	82.3	86.1	74.5
Auto theft	85.7	92.2	87.1	84.4	87.2	77.6
Burglary	84.4	91.7	87.0	80.3	86.2	76.7
Drug offenses	74.4	89.9	86.9	80.1	61.1	53.8
Homicide	80.0	92.9	78.6	71.4	78.6	78.6
Robbery	87.0	94.9	91.7	81.4	89.3	77.8
Intersections only	40.6	52.5	52.3	54.1	12.9	31.5
Assaults	33.8	41.2	39.8	39.8	17.4	31.0
Auto burglary	61.6	71.7	73.8	76.9	25.4	60.4
Auto theft	60.1	74.7	68.3	75.8	24.6	56.9
Burglary	62.3	76.6	48.1	63.6	62.3	61.0
Drug offenses	33.5	46.4	48.3	49.2	3.9	19.5
Homicide	60.0	80.0	40.0	80.0	40.0	60.0
Robbery	66.3	80.7	78.4	77.8	35.7	59.1

Note: Drug offenses include drug possession, trafficking, and narcotics incidents known to law enforcement.

Table A6. Geocoding match rate results for Tampa Police Department by crime type and type of street reference data, 2007-08.

		Street Geocoding				
		Free			Commercial	
Crime events in --	Average	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas
Tampa						
All locations	65.6	70.3	60.5	57.3	69.0	70.8
Without intersections	84.1	90.5	77.5	73.2	88.5	90.6
Assaults	85.1	91.6	78.5	74.5	89.7	91.3
Auto burglary	82.8	89.0	75.9	72.0	87.1	89.8
Auto theft	83.3	90.0	75.9	72.6	87.6	90.4
Burglary	84.8	91.6	77.8	74.1	89.2	91.3
Drug offenses	83.3	89.4	77.7	72.4	87.5	89.6
Homicide	82.4	90.5	69.0	71.4	88.1	92.9
Robbery	82.4	89.2	76.7	69.6	87.4	89.2
Intersections only	3.2	1.9	3.3	3.5	3.4	4.0
Assaults	3.7	2.0	3.1	3.4	5.4	4.7
Auto burglary	1.2	0.8	2.5	2.8	15.1	2.8
Auto theft	4.0	2.5	3.7	4.1	5.4	4.3
Burglary	3.3	1.4	2.8	2.9	4.2	5.4
Drug offenses	3.8	2.2	3.7	3.9	5.4	3.7
Homicide	3.1	0.0	0.0	0.0	7.7	7.7
Robbery	3.3	1.8	2.6	2.6	5.3	4.1

Note: Drug offenses include drug possession, trafficking, and narcotics incidents known to law enforcement.

Appendix B

Table B1. Positional error statistics for street geocoded crime events in all jurisdictions, 2007-08.

		Median positional error (m)					95th percentile positional error (m)				
		Free			Commercial		Free			Commercial	
Crimes in –	Sample	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas
Albuquerque											
All crimes	40,464	46	84	75	67	68	258	331	311	253	370
Assaults	6,700	51	85	79	72	73	300	337	324	267	498
Auto burglary	13,589	48	86	74	67	68	255	356	328	246	352
Auto theft	7,495	51	88	79	72	72	274	363	338	270	389
Burglary	8,643	37	73	64	59	58	228	284	274	222	314
Drug offenses	1,924	83	109	98	91	95	319	317	285	270	1,032
Homicide	50	41	84	84	73	75	236	334	340	265	2,396
Robbery	2,063	60	86	82	69	68	254	307	302	249	405
Arlington											
All crimes	17,050	65	97	88	80	80	289	326	417	270	391
Assaults	744	78	100	94	86	87	2,721	2,688	2,749	2,728	2,749
Auto burglary	8,915	69	102	92	84	83	293	320	417	272	354
Auto theft	1,820	74	96	88	82	84	269	295	349	252	282
Burglary	4,524	47	86	78	69	69	205	259	280	206	233
Drug offenses	243	84	108	101	94	95	2,782	2,801	2,809	2,793	2,769
Homicide	21	75	86	99	63	81	184	285	268	180	183
Robbery	783	84	100	97	84	86	2,722	2,702	2,747	2,735	2,752
Charlotte-Mecklenburg											
All crimes	51,893	61	73	65	61	65	224	281	246	207	848
Assaults	10,277	59	70	62	59	62	220	264	217	188	393
Auto burglary	14,715	68	82	73	70	74	256	310	303	244	2,904
Auto theft	6,505	63	75	67	64	68	226	304	244	202	476
Burglary	13,609	55	68	59	56	59	176	254	205	169	2,564
Drug offenses	3,722	56	63	60	56	58	227	229	212	207	2,144
Homicide	65	56	61	58	59	57	154	174	145	144	154
Robbery	3,000	68	77	70	64	68	276	267	241	213	333

(continued)

Table B1. Positional error statistics for street geocoded crime events in all jurisdictions, 2007-08 (continued).

		Median positional error (m)					95th percentile positional error (m)				
		Free			Commercial		Free			Commercial	
Crimes in –	Sample	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas
Las Vegas											
All crimes	51,730	64	90	69	60	63	273	305	292	246	394
Assaults	1,498	64	90	68	59	62	249	282	276	228	403
Auto burglary	8,931	62	97	69	60	61	292	371	320	279	394
Auto theft	13,613	68	93	71	65	66	293	367	312	278	572
Burglary	17,621	57	86	62	55	56	256	291	276	237	297
Drug offenses	6,105	69	83	75	63	67	273	285	278	229	343
Homicide	112	64	80	75	66	67	216	240	225	214	1,419
Robbery	3,850	79	95	85	71	78	257	307	293	241	1,225
San Diego (County)											
All crimes	16,545	66	83	76	64	62	1,041	10,248	3,641	368	444
Assaults	1,762	64	78	72	63	60	775	3,434	1,068	406	602
Auto burglary	4,619	63	83	72	64	60	497	781	803	322	376
Auto theft	3,390	61	81	71	62	59	521	802	799	289	303
Burglary	4,001	68	82	77	65	64	704	8,031	1,021	314	406
Drug offenses	2,214	75	90	86	71	67	17,943	28,590	20,331	873	2,682
Homicide	17	99	92	60	69	61	26,177	31,994	27,890	10,231	10,269
Robbery	542	75	87	86	65	64	13,864	25,493	44,730	335	675
Tampa											
All crimes	16,193	64	64	64	62	60	220	274	258	220	314
Assaults	5,034	63	63	64	62	60	217	253	258	216	327
Auto burglary	2,529	62	63	63	61	59	217	274	253	220	273
Auto theft	1,473	66	66	64	64	60	219	294	258	218	283
Burglary	3,327	64	63	65	63	60	223	285	258	208	325
Drug offenses	3,057	65	66	65	65	59	231	274	258	259	314
Homicide	19	48	37	47	48	38	16,860	133	16,773	156	119
Robbery	754	63	63	65	61	62	218	274	258	238	294

Note: Drug offenses include drug possession, trafficking, and narcotics known to law enforcement. The sample size represents the total number of parcel geocoded crime incidents used to reference positional accuracy.

Appendix C

Table C1. Measures of positional error (in meters) of geocoding results by crime type and type of street reference data for all jurisdictions, 2007-2008.

	Free			Commercial	
Crimes and measures	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas
All crimes					
Minimum	1.28	0.37	2.02	0.82	0.34
Maximum	88,444.71	88,456.92	88,447.22	88,442.56	88,443.86
Median	60.17	81.26	70.26	64.14	65.38
68th Percentile	93.40	117.40	104.46	93.69	97.04
90th Percentile	186.11	235.70	213.11	178.82	212.22
95th Percentile	263.41	322.24	297.34	244.41	412.99
99th Percentile	2,731.87	2,544.49	2,669.53	596.34	12,282.70
Assaults					
Minimum	3.09	0.46	3.68	4.51	4.57
Maximum	80,959.26	74,226.57	74,226.02	69,762.89	69,762.89
Median	59.43	73.46	67.12	62.50	63.87
68th Percentile	87.43	104.61	95.10	88.81	92.14
90th Percentile	174.78	211.72	189.13	171.23	203.28
95th Percentile	255.26	302.87	275.75	236.13	402.73
99th Percentile	2,546.47	1,695.64	1,572.22	553.90	10,866.08
Auto Burglary					
Minimum	2.74	0.37	2.02	0.82	3.13
Maximum	81,289.05	77,548.70	77,561.38	77,555.34	77,591.97
Median	62.59	88.12	75.52	68.87	70.16
68th Percentile	99.18	125.23	110.64	99.67	101.80
90th Percentile	186.81	248.19	228.81	186.54	220.86
95th Percentile	274.01	345.05	330.14	256.71	446.82
99th Percentile	2,444.82	2,629.07	2,740.53	857.98	11,708.74
Auto Theft					
Minimum	3.34	0.44	3.02	1.56	3.70
Maximum	81,289.05	77,882.34	77,913.02	71,265.85	71,265.24
Median	63.06	85.48	72.57	66.61	67.65
68th Percentile	101.23	125.43	110.64	100.30	101.31
90th Percentile	205.05	252.21	233.74	193.40	223.42
95th Percentile	283.91	358.71	313.31	256.89	456.20
99th Percentile	2,617.76	1,470.56	2,646.71	596.34	12,581.75

(continued)

Table C1. Measures of positional error (in meters) of geocoding results by crime type and type of street reference data for all jurisdictions, 2007-2008 (continued).

	Free			Commercial	
Crimes and measures	Local Centerlines	StreetMap USA	TIGER	NAVTEQ	Tele Atlas
Burglary					
Minimum	1.28	0.66	2.02	1.56	0.34
Maximum	80,830.81	83,275.54	76,158.36	76,181.50	76,180.66
Median	53.54	76.48	64.15	58.25	59.11
68th Percentile	81.18	109.14	94.02	84.05	87.47
90th Percentile	168.58	221.45	197.10	162.10	194.27
95th Percentile	233.88	292.71	272.44	219.89	325.10
99th Percentile	921.52	878.37	869.00	408.59	13,105.67
Drug Offense					
Minimum	6.12	0.74	6.22	2.43	4.46
Maximum	88,444.71	88,456.92	88,447.22	88,442.56	88,443.86
Median	66.22	77.70	71.01	64.44	65.99
68th Percentile	101.24	116.56	106.72	97.18	101.03
90th Percentile	204.42	234.67	215.00	189.75	227.75
95th Percentile	303.83	331.93	302.70	267.52	697.01
99th Percentile	21,470.79	12,266.49	10,571.90	1,687.49	13,350.61
Homicide					
Minimum	7.02	3.63	7.11	13.04	13.73
Maximum	33,477.12	44,689.58	37,390.49	50,361.93	50,344.09
Median	59.07	73.71	69.57	65.55	65.93
68th Percentile	97.58	106.55	103.34	94.27	97.71
90th Percentile	175.74	197.49	186.21	159.44	182.30
95th Percentile	241.05	284.36	275.08	216.45	991.74
99th Percentile	15,236.57	8,639.36	18,708.29	2,954.98	10,216.21
Robbery					
Minimum	1.52	1.42	4.07	3.91	5.68
Maximum	81,289.05	74,226.57	74,226.02	57,470.26	57,473.95
Median	71.34	84.85	78.55	67.85	71.00
68th Percentile	105.39	124.88	112.95	100.01	103.00
90th Percentile	195.18	244.77	225.86	182.49	218.85
95th Percentile	279.16	318.45	297.25	243.60	598.36
99th Percentile	3,238.80	2,816.77	2,868.88	1,092.68	11,669.27

Note: Drug offenses include drug possession, trafficking, and narcotics incidents known to law enforcement.